Министерство образования и науки РФ
АМУРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Факультет математики и информатики

Н.Н. Двоерядкина, А.Н. Киселева, Т.А. Юрьева

МНОГОМЕРНЫЕ МАТЕМАТИЧЕСКИЕ МЕТОДЫ МОДЕЛИРОВАНИЯ СОЦИАЛЬНЫХ ПРОЦЕССОВ:

Учебно-методическое пособие

Благовещенск
2012

Печатается по решению
редакционно-издательского совета
факультета математики и информатики
Амурского государственного
университета

Двоерядкина Н.Н., Киселева А.Н., ЮрьеваТ.А.

Многомерные математические методы моделирования социальных процессов. Учебно-методическое пособие. – Благовещенск: Амурский гос. ун-т, 2012.

В пособии рассматриваются основные многомерные методы математической статистики, используемые для моделирования социальных процессов. Приведены краткие теоретические сведения и типовые задачи, наиболее часто встречающиеся в экспериментальных социологических исследованиях. Содержатся задания для организации самостоятельной работы студентов. Предназначено для студентов специальности 040201.65 и направления подготовки 040200.62.

Рецензент: Бурдуковская Е.А. кандидат педагогических наук, доцент кафедры педагогики и психологии АмГУ

© Амурский государственный университет, 2012

Содержание

ВВЕДЕНИЕ 4
1 МНОГОМЕРНЫЙ АНАЛИЗ ДЛЯ ИССЛЕДОВАНИЯ СОЦИАЛЬНЫХ ПРОЦЕССОВ. КЛАССИФИКАЦИЯ МНОГОМЕРНЫХ МЕТОДОВ..... 5
2 КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ..... 8
2.1 Модель парной линейной регрессии 8
2.2 Виды коэффициентов корреляции 12
2.3 Оценка качества уравнения регрессии 16
2.4 Нелинейная регрессия 19
2.5 Модель множественной регрессии. Спецификация модели 21
3 КЛАСТЕРНЫЙ АНАЛИЗ..... 27
3.1 Постановка задачи 27
3.2 Иерархические кластерные структуры 30
3.3 Метод k-средних..... 36
4 ФАКТОРНЫЙ АНАЛИЗ, И ЕГО ИСПОЛЬЗОВАНИЕ В ИССЛЕДОВАНИИ СВЯЗИ 46
4.1 Постановка задачи 46
4.2 Моделирование значений наблюдаемых переменных на основе выделенных латентных факторов..... 53
4.3 Методы главных компонент и главных факторов, вращение факторов 56
5 ДИСКРИМИНАНТНЫЙ АНАЛИЗ 68
5.1 Постановка задачи 68
5.2 Построение дискриминантных функций 69
5.3 Построение классификации объектов 74

ВВЕДЕНИЕ

Современное исследование общественных и социально-экономических процессов опирается в значительной степени на использование математических методов анализа. Это требует от специалиста, способного к проведению исследований, достаточно свободного владения математическим аппаратом изучения статистических данных. Поэтому важное место в подготовке социолога занимает учебный курс «Математические методы моделирования социальных процессов».

Наиболее сложными для исполнения являются многомерные математические методы. В этой связи решение социологических задач, связанных с применением многомерных методов, осуществляется с помощью специализированных компьютерных программ (SPSS, Statistica и др.). Тем не менее, для адекватного, осознанного применения многомерных методов и объективной интерпретации результатов необходимо глубокое понимание сути и теоретических основ осуществляемых процедур. Именно поэтому основное содержание учебно-методического пособия посвящено изложению «ручных» способов осуществления многомерных методов.

Пособие начинается с краткого изложения теоретического материала, который подкрепляется конкретными примерами из разных областей социологии, демонстрирующими пути и способы применения тех методов теоретико-вероятностного описания и проверки статистических гипотез. В пособие приведены задания для самостоятельной работы студентов. При работе над пособием читатели должны опираться на систему базовых математических знаний, приобретенных при изучении высшей математики, теории вероятностей и математической статистики, и понимать качественный смысл тех количественных преобразований в области социологии, которые они осуществляют с помощью математического инструментария.

Настоящее пособие будет полезно не только студентам, но и преподавателям, аспирантам и соискателям в области социологических наук.

1 МНОГОМЕРНЫЙ АНАЛИЗ ДЛЯ ИССЛЕДОВАНИЯ СОЦИАЛЬНЫХ ПРОЦЕССОВ. КЛАССИФИКАЦИЯ МНОГОМЕРНЫХ МЕТОДОВ

Социально-экономические процессы и явления зависят от большого числа параметров, их характеризующих, что обусловливает трудности, связанные с выявлением структуры взаимосвязей этих параметров. В подобных ситуациях, т. е. когда решения принимаются на основании анализа стохастической, неполной информации, использование методов многомерного статистического анализа является не только оправданным, но и существенно необходимым.

Многомерные статистические методы среди множества возможных вероятностно-статистических моделей позволяют обоснованно выбрать ту, которая наилучшим образом соответствует исходным статистическим данным, характеризующим реальное поведение исследуемой совокупности объектов, оценить надежность и точность выводов, сделанных на основании ограниченного статистического материала.

К области приложения многомерных статистических методов могут быть отнесены задачи, связанные с исследованием поведения индивидуума, семьи или другой социально-экономической или производственной единицы, как представителя большой совокупности объектов.

Выделяют три центральные задачи, решаемые с помощью многомерных методов.

1. Статистическое исследование структуры и характера взаимосвязей, существующих между анализируемыми количественными переменными. При этом под переменными понимаются как регистрируемые на объектах признаки, так и время *t*.

2. Разработка статистических методов классификации объектов и признаков.

3. Снижение размерности исследуемого признакового пространства с целью лаконичного объяснения природы анализируемых многомерных данных.

Возможность лаконичного описания анализируемых многомерных данных основана на априорном допущении, в соответствии с которым существует

небольшое число признаков, с помощью которых могут быть достаточно точно описаны как сами наблюдаемые переменные анализируемых объектов, так и определяемые этими переменными свойства (характеристики) анализируемой совокупности. При этом упомянутые признаки могут находиться среди исходных признаков, а могут быть латентными, т.е. непосредственно статистически не наблюдаемыми, но восстанавливаемыми по исходным данным.

Эти задачи не исчерпывают всех возможностей многомерных статистических методов, но в настоящий момент являются наиболее распространенными.

В соответствии с задачами в структуре многомерных статистических методов выделяют методы снижения размерности, методы исследования зависимостей, методы классификации.

Классификация многомерных методов представлена на рисунке 1.

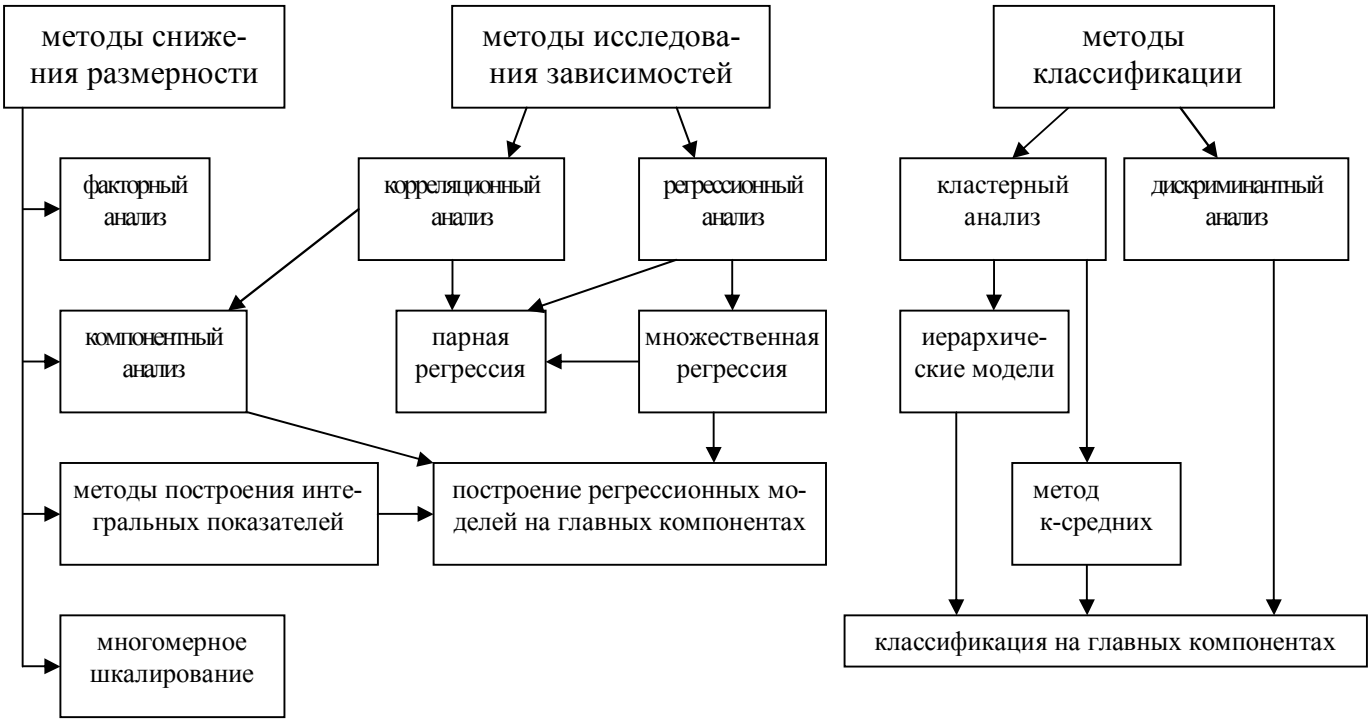

Рис. 1

Работая с многомерными статистическими методами важно, чтобы переменные изменялись в сравнимых шкалах. Из неоднородности единиц измерения вытекает невозможность обоснованного выражения значений различных показателей в одном масштабе. Чтобы устранить неоднородность измерения

исходных данных, все их значения предварительно нормируются, т.е. выражаются через отношение этих значений к некоторой величине, отражающей определенные свойства данного показателя. Нормирование исходных данных иногда проводится посредством деления исходных величин на среднеквадратичное отклонение соответствующих показателей. Другой способ сводится к вычислению, так называемого, стандартизованного вклада. Его еще называют *Z*-вкладом.

Z - вклад показывает сколько стандартных отклонений отделяет данное наблюдение от среднего значения:

$$Z_i = \frac{x_i - \bar{x}}{\sigma_i},$$

где x_i - значение данного наблюдения, $\bar{x}$ - среднее, σ_i -стандартное отклонение.

Среднее для *Z*-вкладов является нулевым и стандартное отклонение равно единице.

Стандартизация позволяет сравнивать наблюдения из различных распределений.

Методы нормирования означают признание всех признаков равноценными с точки зрения выяснения сходства рассматриваемых объектов. Признание равноценности различных показателей кажется оправданным отнюдь не всегда. Было бы желательным наряду с нормированием придать каждому из показателей вес, отражающий его значимость в ходе установления сходств и различий объектов.

В этой ситуации приходится прибегать к способу определения весов отдельных показателей – опросу экспертов.

Экспертные оценки дают известное основание для определения важности индикаторов, входящих в ту или иную группу показателей.

Довольно часто при решении подобных задач используют не один, а два расчета: первый, в котором все признаки считаются равнозначными, второй,

где им придаются различные веса в соответствии со средними значениями экспертных оценок.

2 КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

2.1 Модель парной линейной регрессии

Связь, при которой каждому значению независимой переменной (аргумента x) ставится в соответствие определенное значение зависимой переменной (функции y) называется функциональной.

Функция может быть однозначной и многозначной. В случае однозначной функции, каждой паре значений x и y соответствует некоторая точка плоскости. Множество этих точек на плоскости представляет собой графическое изображение функциональной связи, или график функции $y=f(x)$.

Однако в природе существуют не только функциональные связи.

Случается, что с изменением одной переменной происходит изменение распределения другой переменной. Связь этих переменных называется статистической. Если оказывается, что с изменением одной переменной изменяется среднее значение другой, то говорят, что между этими переменными существует корреляционная связь.

Изучением приемов, с помощью которых исследуются и обобщаются корреляционные связи, занимается корреляционный анализ. А количественное описание связи осуществляется на основе регрессионного анализа.

Очень часто необходимо дать количественное описание взаимосвязей между переменными. Переменные, с которыми приходится работать являются случайными величинами. Можно выделить три вида случайных величин:

1. Дискретные;
2. Непрерывные, нормально распределенные;
3. Непрерывные, не подчиняющиеся нормальному распределению.

В связи с тем, что многие реальные величины хотя и дискретны, но имеют очень большое число возможных значений, их считают непрерывными (доход населения, курсы валют, и т.д.).

Наиболее распространенным является нормальное распределение непрерывных случайных величин, которое характеризуется тем, что большое скопление значений находится вблизи среднего значения, а при удалении от среднего количество наблюдений уменьшается. Нормальное распределение характеризуется кривой Гаусса.

Для оценки связи между показателями при большом числе наблюдений чаще всего используют линейные зависимости, поэтому наибольший интерес представляет линейная регрессия.

В зависимости от количества переменных (факторов), включенных в модель, различают парную и множественную регрессии.

Парная регрессия – уравнение связи двух переменных y и x:

$$y = \hat{f}(x),$$

где y – зависимая переменная (результативный признак);

x – независимая переменная (признак – фактор).

Например, требуется определить зависимость между ростом жены и мужа. Для примера рассмотрим 100 супружеских пар (рис. 2). На плоскости дана прямоугольная система координат, по оси x откладывается рост мужа, по оси y – рост жены. Точкой на плоскости отмечается каждая супружеская пара. Полученное графическое изображение называется корреляционным полем.

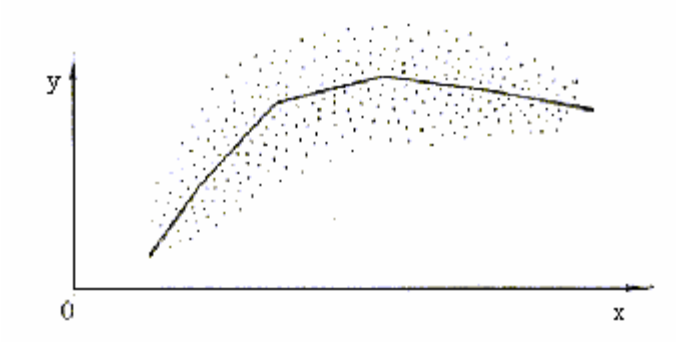

Рис. 2

В нашем случае должно быть 100 точек, которые как-то заполняют плоскость этого корреляционного поля. Разобьем все значения переменной x на несколько интервалов. Для каждого класс-интервала x отбираем все соответст-

вующие ему точки. Находим их среднее значение $\bar{y}$. Эту точку наносим на график, обозначая ее крестиком, чтобы выделить среди прочих. Соединяем последовательно отрезками прямых все точки отмеченные крестиком. Полученная ломаная линия показывает изменение среднего значения роста жены с изменением роста мужа от одного класс-интервала к другому. Эта линия называется эмпирической линией регрессии.

Если выбрать 100 других пар, то получится несколько иная эмпирическая линия регрессии. Если уменьшить величину класс-интервала, то линия покажет увеличение числа звеньев, сохранив в целом контур. Можно убедиться, что все эмпирические линии регрессии каких-либо двух переменных всегда лежат около некоторой плавной линии, называемой теоретической линией регрессии, или просто линией регрессии. Ее уравнение называется уравнением регрессии. Если мы рассматриваем изменение среднего y от x, то получится уравнение регрессии y на x: $\bar{y}_x = \varphi(x)$.

Если рассматриваем изменение среднего $\bar{x}$ от y, то уравнение регрессии x на y: $\bar{x}_y = \varphi(y)$.

При $\varphi(x) = ax + b$ говорят о линейной регрессии y на x, т.е. $\bar{y}_x = ax + b$. Аналогично можно ввести уравнение регрессии x на y: $\bar{x}_y = cy + d$.

Коэффициенты теоретической линии регрессии находят по методу наименьших квадратов: ищут эту линию при том условии, чтобы сумма квадратов расстояний между ординатами эмпирической и теоретической линий регрессии была минимальной. Иначе говоря, теоретическая линия регрессии должна иметь наиближайшее расположение ко всем точкам эмпирической линии регрессии.

Если мы обозначили ординату теоретической линии регрессии y_m, эмпирической – $y_э$, то надо найти минимум величины:

$$f = \sum (y_э - y_m)^2 = \sum (y_э - a - bx)^2.$$

Это означает, что

$$\begin{cases} \dfrac{\partial f}{\partial a} = 2\sum (y - a - bx)(-1) = 0 \\ \dfrac{\partial f}{\partial b} = 2\sum (y - a - bx)(-x) = 0 \end{cases}$$

Получаем нормальные уравнения для определения коэффициентов линии регрессии:

$$\begin{cases} \sum [y - (a + bx)] = 0; \\ \sum [y - (a + bx)]x = 0. \end{cases} \text{ или } \begin{cases} \sum y = na + b\sum x; \\ \sum xy = a\sum x + b\sum x^2. \end{cases} \text{ или } \begin{cases} \bar{y} = a + b\bar{x} \\ \overline{xy} = a\bar{x} + b\overline{(x^2)} \end{cases}$$

Найденный коэффициент b называется коэффициентом регрессии y на x и обозначается ρ_{yx}. Аналогично определяется коэффициент регрессии x на y и обозначается ρ_{xy}.

Коэффициент корреляции является мерой наличия связи между переменными и обозначается *r*.

Геометрическая интерпретация коэффициента корреляции представлена на рисунке 3.

OP – это линия регрессии *y* на *x*;

OQ – линия регрессии *x* на *y*; $\rho_{yx} = tg\alpha$; $\rho_{xy} = tg\beta$ $r_{xy} = \sqrt{tg\alpha tg\beta}$.

Если корреляции нет, то или линия OP, или OQ или обе вместе совпадают с осями координат, так как: α =0 или β =0 и, следовательно, *r* = 0.

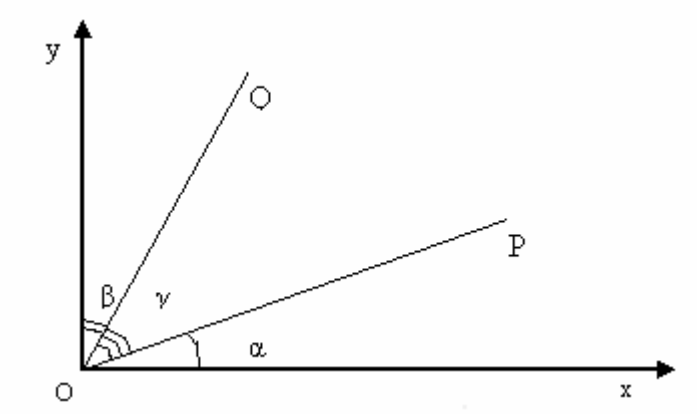

Рис. 3

Если корреляционная связь переходит в функциональную, то обе линии

регрессии совпадают. Тогда $\alpha+\beta=90^\circ$, $r=\sqrt{tg\alpha tg\beta}=\sqrt{tg\alpha tg\left(90^\circ-\alpha\right)}=\sqrt{tg\alpha ctg\alpha}=1$ т.е. коэффициент корреляции равен 1.

Чем теснее связь между переменными, тем меньше угол между обеими линиями регрессии.

Рассмотренный коэффициент корреляции измеряет линейную связь между двумя количественными переменными. Этим, однако, не исчерпывается все возможное многообразие связей в социологии.

Во-первых, переменные могут иметь криволинейную регрессию: линия регрессии может быть параболой, кубической параболой, экспонентой и т.п. В каждом случае надо находить пути измерения связи между данными переменными.

Во-вторых, возможно наличие связи между более чем двумя переменными. Это проблема множественной корреляции, или многофакторного корреляционного анализа.

В-третьих, возможно существование связи между не только количественными переменными. В этом случае в статистике и социологии используются специальные показатели связи.

2.2 Виды коэффициентов корреляции

Задача корреляционного анализа сводится к установлению направления и формы между варьирующими признаками, измерению тесноты, и, наконец, к проверке значимости коэффициентов корреляции.

Переменные *x* и *y* могут быть измерены в разных шкалах. Это обстоятельство определяет выбор соответствующего коэффициента корреляции (таблица 1).

Таблица 1

Тип шкалы		Мера связи
Переменная *x*	Переменная *y*	
Интервальная или отношений	Интервальная или отношений	Коэффициент Пирсона r_{xy}
Ранговая, интервальная или отношений	Ранговая, интервальная или отношений	Коэффициент Спирмена g_{xy}

Ранговая	Ранговая	Коэффициент τ Кендалла
Дихотомическая	Дихотомическая	Коэффициент ассоциации φ
Дихотомическая	Ранговая	Рангово-бисериальный $R_{гв}$
Дихотомическая	Интервальная или отношений	Бисериальный $R_{бис}$

Все коэффициенты по абсолютной величине не могут превосходить 1.

Рассмотрим некоторые виды коэффициентов корреляции.

а) Коэффициент корреляции Пирсона вычисляется по формуле:

$$r_{xy} = \frac{\sum_i (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 * \sum_i (y_i - \bar{y})^2}} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\left(\overline{x^2} - \left(\bar{x}\right)^2\right)} \cdot \sqrt{\left(\overline{y^2} - \left(\bar{y}\right)^2\right)}}$$

где x_i – значения, переменных принимаемые переменной x,

y_i - значения, переменных принимаемые переменной y,

$\bar{x}$ – средняя по x,

$\bar{y}$ - средняя по y.

Оценка значимости осуществляется при числе степеней свободы $k=n-2$.

б) Коэффициент корреляции рангов Спирмена вычисляется по формуле:

$$\rho = 1 - \frac{6 \cdot \sum_i (d_i)^2}{n \cdot (n^2 - 1)}$$

где n – количество ранжируемых признаков

d_i – разность между рангами по двум переменным для каждого испытуемого.

При наличии одинаковых рангов в числитель добавляются поправки на одинаковые ранги:

$$D_1 = \frac{n^3 - n}{12} \qquad D_2 = \frac{k^3 - k}{12},$$

где n – число одинаковых рангов в первом столбце,

k – число одинаковых рангов во втором столбце.

По каждой группе одинаковых рангов вводится своя поправка.

Критически значения определяются при уровне значимости равном числу

значений признака, по таблице критических значений ρ Спирмена.

в) Коэффициент ассоциации φ вычисляется по формуле:

$$\varphi = \frac{p_{xy} - p_x \cdot p_y}{\sqrt{p_x \cdot (1 - p_x) \cdot p_y \cdot (1 - p_y)}},$$

где p_x – частота или доля признака, имеющего 1 по x,

$(1-p_x)$ - частота или доля признака, имеющего 0 по x,

p_y - частота или доля признака, имеющего 1 по y,

$(1-p_y)$ - частота или доля признака, имеющего 0 по y,

p_{xy} - частота или доля признака, имеющих 1 и по x и по y.

г) Коэффициент корреляции τ *Кендалла* вычисляется по формуле:

$$\tau = 1 - \frac{4 \cdot Q}{N \cdot (N-1)},$$

где Q – число инверсий (подсчет инверсий осуществляется суммированием числа рангов второго признака меньше каждого из рангов второго признака, при условии, что ранги первого признака упорядочены по возрастанию).

д) Бисериальный коэффициент корреляции вычисляется по формуле:

$$R_{бис\ эмп} = \frac{\bar{x}_1 - \bar{x}_0}{\sigma_y} * \sqrt{\frac{n_1 * n_0}{N * (N-1)}},$$

где $\bar{x}_1$ - среднее по тем элементам переменной *y*, которым соответствует признак 1 в переменной *x*,

$\bar{x}_0$ - среднее по тем элементам переменной *y*, которым соответствует признак 0 в переменной *x*,

n_1 – число единиц в переменной *x*,

n_0 – число нулей в переменной *x*,

$N=n_1+n_0$,

σ_y - среднее квадратическое отклонение переменной *y*.

е) Рангово-бисернальный коэффициент корреляции вычисляется по формуле:

$$R_{rb\ эмп} = \frac{(\bar{x}_1 - \bar{x}_0)^2}{N},$$

где $\bar{x}_1$ - средний ранг по тем элементам переменной *y*, которым соответствует признак 1 в переменной *x*;

$\bar{x}_0$ - средний ранг по тем элементам переменной *y*, которым соответствует признак 0 в переменной *x*;

N – количество элементов в переменной *x*.

Для измерения нелинейной корреляционной связи между признаками используется *корреляционное отношение*.

Показатели корреляционного отношения вычисляется по формулам:

$$h_{yx} = \sqrt{\frac{\sum f_x (\bar{y}_x - \bar{y})^2}{\sum f_y (y_i - \bar{y})^2}},$$

$$h_{xy} = \sqrt{\frac{\sum f_y (\bar{x}_y - \bar{x})^2}{\sum f_x (x_i - \bar{x})^2}},$$

где $\bar{x}$ и $\bar{y}$ общие, а $\bar{x}_y$ и $\bar{y}_x$ - групповые средние арифметические, f_y и f_x частоты рядов *X* и *У*.

Для вычисления корреляционного отношения h_{xy} (*У* по *X*) или h_{xy} (*X* по *У*) необходимо:

1) расположить по порядку исходные данные по *X* от меньшей к большей, при этом сохранив значения соответствующих величин *У* по отношению к *X*.

2) Определить частоты переменной $X(f_x)$.

3) Подсчитать арифметическое средние по переменной *У* для соответствующей частоты $(\bar{y}_x)$;

4) Расположить по порядку исходные данные по *У* от меньшей величине к большей, при этом сохранив значения соответствующих величин *X* по отношению к *У*;

5) Определить частоты переменной $У(f_y)$;

6) Подсчитать арифметические средние по переменной X для соответствующей частоты $(\bar{x}_y)$;

7) Определить $\bar{x}$ и $\bar{y}$;

8) Произвести расчет по формулам;

9) Определить значимость полученных показателей.

Расчет уровней значимости коэффициентов корреляции.

При оценке значимости коэффициентов корреляции рассматриваются гипотезы:

H_0: коэффициент корреляции между признаками статистически значимо не отличается от нуля;

H_1: коэффициент корреляции между признаками статистически значимо отличается от нуля.

Все коэффициенты корреляции, не имеющие стандартных таблиц для нахождения критических значений, оценивают с помощью t – критерия Стьюдента по формуле:

$$t_{эмп} = |r_{эмп}| * \sqrt{\frac{n-2}{1-r^2{}_{эмп}}},$$

где $r_{эмп}$ – соответствующий коэффициент корреляции (корреляционное отношение),

n – число коррелируемых значений.

Критические значения $t_{кр}$ определяется по таблице значений для t – критерия Стьюдента. Число степеней свободы равно $k=n-2$.

2.3 Оценка качества уравнения регрессии

Из дисперсионного анализа известно, что

$$\sum\left(y-\bar{y}\right)^2 = \sum\left(\hat{y}_x-\bar{y}\right)^2 + \sum\left(y-\hat{y}_x\right)^2,$$

где $\sum\left(y-\bar{y}\right)^2$ – общая сумма квадратов отклонений (TSS); $\sum\left(\hat{y}_x-\bar{y}\right)^2$ – сумма квадратов отклонений, обусловленная регрессией (RSS); $\sum\left(y-\hat{y}_x\right)^2$ – остаточная сумма квадратов отклонений (ESS).

Долю дисперсии, объясняемую регрессией, в общей дисперсии результативного признака y характеризует *коэффициент детерминации:*

$$R^2 = \frac{\sum\left(\hat{y}_x - \bar{y}\right)^2}{\sum\left(y - \bar{y}\right)^2}.$$

Для парной линейной регрессии коэффициент детерминации равен квадрату коэффициента корреляции.

Чем ближе значение коэффициента детерминации к 1 тем выше качество регрессии.

Для оценки качества уравнения регрессии используют статистику Фишера (F-тест). В этом случае проверяется гипотеза H_0 о статистической незначимости уравнения регрессии и показателя тесноты связи.

H_0: уравнение регрессии статистически незначимо.

H_1: уравнение регрессии статистически значимо.

Для проверки гипотез выполняется сравнение фактического (наблюдаемого) $F_{факт}$ и критического (табличного) $F_{табл}$ значений F-критерия Фишера.

Фактическое значение критерия определяется по формуле:

$$F_{факт} = \frac{\sum\left(\hat{y} - \bar{y}\right)^2 / m}{\sum\left(y - \hat{y}\right)^2 / \left(n - m - 1\right)} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m},$$

где n – число единиц совокупности; m – число параметров при совокупности x.

$F_{табл}$ – это максимально возможное значение критерия под влиянием случайных факторов при данных степенях свободы и уровне значимости α. Уровень значимости α – вероятность отвергнуть правильную гипотезу. Обычно α принимается равной 0,05 или 0,01.

Если $F_{табл} < F_{факт}$, то принимается гипотеза H_1, о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость и надежность. Если $F_{табл} > F_{факт}$, то принимается гипотеза H_0 и признается статистическая незначимость, ненадежность уравнения регрессии.

Для оценки существенности отдельных параметров уравнения определяется их стандартная ошибка m_b и m_a. В этом случае выполняется проверка двух вариантов гипотез.

1 вариант: проверка значимости коэффициента b

H_o: b_i=0 (коэффициент регрессии равен нулю)

H_1: $b_i \neq 0$ (коэффициент регрессии неравен нулю)

$$m_b = \sqrt{\frac{\sum (y - \hat{y}_x)^2 / (n - m - 1)}{\sum (x - \bar{x})^2}} = \sqrt{\frac{S^2}{\sum (x - \bar{x})^2}},$$

где S^2 – остаточная дисперсия на одну степень свободы.

Величина $t_b = \frac{b}{m_b}$ имеет t-распределение Стьюдента с $n - m - 1$ степенями свободы, которое затем сравнивается с табличным значением при определенном уровне значимости α. Доверительный интервал для коэффициента регрессии определяется как $b \pm t_{\alpha, n-2} \cdot m_b$.

2 вариант: проверка значимости коэффициента a

H_o: a=0 (коэффициент регрессии равен нулю)

H_a: a $\neq$0 (коэффициент регрессии неравен нулю)

Стандартная ошибка параметра a определяется по формуле

$$m_a = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{n - m - 1} \cdot \frac{\sum x^2}{n \cdot \sum (x - \bar{x})^2}} = S \sqrt{\frac{\sum x^2}{n \cdot \sum (x - \bar{x})^2}}.$$

Процедура оценивания значимости данного параметра ничем не отличается от процедуры оценки значимости коэффициента регрессии t_b.

Доверительный интервал для прогноза

Точечный прогноз $\hat{y}_x$ значения признака часто бывает нереальным и дополняется расчетом стандартной ошибки $m_{\hat{y}_x}$ и соответственно интервальной оценкой прогнозного значения (y^*): $\hat{y}_x - m_{\hat{y}_x} \le y^* \le \hat{y}_x + m_{\hat{y}_x}$.

Средняя ошибка прогноза вычисляется по формуле: $m_{\hat{y}_{прогн}} = \sigma_{ост} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_{прогн} - \bar{x})^2}{\sum (x - \bar{x})^2}}$, где $\sigma_{ост} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - m - 1}} = \sqrt{\frac{161{,}402}{10}} = 4{,}017$,

Строим доверительный интервал с заданной доверительной вероятностью $(\gamma_{\hat{y}_{\min}}, \gamma_{\hat{y}_{\max}})$: $(\hat{y}_{прогн} - t_{табл} \cdot m_{\hat{y}_{прогн}}, \hat{y}_{прогн} + t_{табл} \cdot m_{\hat{y}_{прогн}})$.

2.4 Нелинейная регрессия

В силу многообразия и сложности социально-экономических процессов многие зависимости не являются линейными и, следовательно, требуют моделирования нелинейными уравнениями.

Для случая парной регрессии подбор модели обычно осуществляется по виду расположения наблюдаемых точек на корреляционном поле, с учетом фактов известных из теории. В случае если зависимость может быть описана несколькими функциями, необходимо выбрать ту из них, которая обладает наилучшим качеством. Но следует помнить, что чем сложнее модель, тем менее интерпретируемы ее параметры.

При визуализации данных на корреляционном поле возможны следующие результаты (рис. 4).

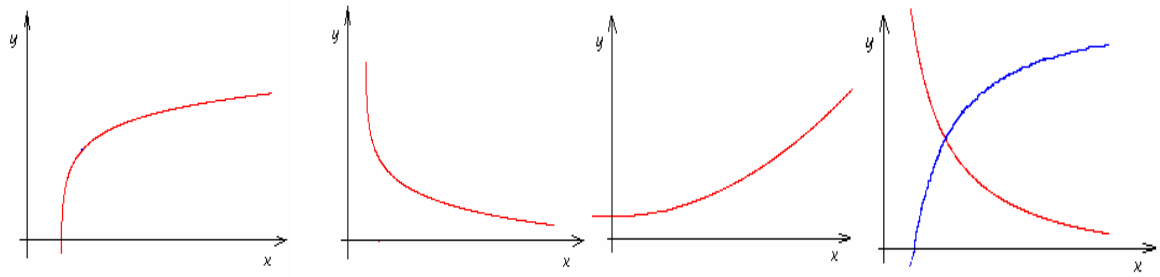

Рис. 4

Рассмотрим наиболее часто встречающиеся модели. Для простоты изложения и возможности графической интерпретации ограничимся моделями

парной нелинейной регрессии.

1. Логарифмические модели.

- $\ln y = \beta_0 + \beta \cdot \ln x + \varepsilon$ - *двойная логарифмическая модель*

коэффициент β в данной модели определяет эластичность переменной Y по переменной X, т.е. процентное изменение Y для данного процентного изменения X.

- $\ln y = a + b \cdot x + \varepsilon$ - *лог-линейная модель*, используется, например, при исследовании зависимости прироста объема выпуска от относительного увеличения затрат ресурса.

Коэффициент b в данной модели имеет смысл темпа прироста переменной y по переменной x , т.е. характеризует отношение относительного изменения y к абсолютному изменению x. Умножив b на 100%, получим процентный темп прироста переменной y.

- $y = a + b \cdot \ln x + \varepsilon$ - *линейно-логарифмическая модель*, используется, например, когда необходимо исследовать влияние процентного изменения независимой переменной на абсолютное изменение зависимой переменной.

Коэффициент b определяет изменение переменной y вследствии единичного относительного прироста x.

2. Гиперболическая модель

$y = a + b \cdot \frac{1}{x} + \varepsilon$ - применяется в тех случаях, когда неограниченное увеличение значений объясняющей переменной x асимптотически приближает зависимую переменную y к некоторому пределу a.

3. Полиномиальная модель.

$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k$

При $k = 2$ квадратичная функция может отражать зависимость между расходами на рекламу (x) и прибылью (y), но большое применение имеет при анализе временных рядов; при $k = 3$ кубическая функция моделирует зависимость общих издержек (y) от объема выпуска (x).

4. Степенная модель.

$y = ax^b$ - отражает, например, зависимость спроса y на благо от его цены или от дохода x. Данная регрессия, путем математических преобразований сводится к двойной логарифмической модели. Коэффициент b является коэффициентом эластичности переменной y по переменной x.

5. Показательная модель.

$y = a \cdot e^{b \cdot x}$ - используется чаще всего в той ситуации, когда анализируется изменение переменной y с постоянным темпом прироста во времени. Переменная x заменяется на t, а модель используется при анализе временных рядов.

Коэффициент b показывает постоянный темп прироста переменной y во времени.

Показательная модель, путем логарифмирования сводится к лог-линейной модели.

Метод наименьших квадратов трудно применим для нахождения эмпирических формул, отличных от многочленов. Однако во многих случаях путем замены переменных можно привести эти зависимости к линейным.

Сведение искомой зависимости к линейной называется линеаризацией функции.

Например, гиперболическая модель сводится к линейной с помощью замены $X = \frac{1}{x}$; лог-линейная – $У = \ln y$; линейно-логарифмическая – $X = \ln x$; двойная логарифмическая модель сводится к линейной с помощью замены $У = \ln y$, $X = \ln x$.

2.5 Модель множественной регрессии. Спецификация модели

На любой статистический показатель действует чаще всего не один, а несколько факторов. Все их следует включить в модель, т.е. построить уравнение множественной регрессии. Рассмотрим самую употребляемую и наиболее простую из моделей множественной регрессии - модель множественной линейной регрессии вида:

$y = \alpha + \beta x_1 + \beta x_2 + \ldots + \beta x_n + \varepsilon,$

где y - зависимая переменная,

$x_1, x_2,, ..., x_n$ - независимые переменные,

ε -случайная величина.

Параметры модели множественной регрессии определяются с помощью метода наименьших квадратов.

Под спецификацией модели подразумевается построение такой модели, которая удобна для проведения анализа. Правильно специфицированное уравнение регрессии верно отражает соотношения между экономическими показателями, участвующими в модели.

Неправильный выбор функциональной формы уравнения или набора объясняющих переменных называется ошибками спецификации. Рассмотрим основные типы ошибок спецификации.

1. В модели отсутствуют значимые переменные.
2. В модель включены несущественные переменные.
3. Выбор неправильной функциональной формы.

Ошибки спецификации допускаются в основном из-за поверхностных знаний об исследуемом объекте либо из-за погрешностей сбора и обработки статистических данных при построении уравнения регрессии. Важно уметь обнаружить и исправить эти ошибки.

Правильный выбор функциональной формы модели осуществляется при согласовании эмпирических данных с теорией. Однако чем сложнее форма модели, тем менее интерпретируемы ее параметры.

Рассмотрим способ обнаружения несущественной переменной в модели на примере.

Пусть теоретическая модель имеет вид: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

Исследователь заменяет ее более сложной моделью, добавляя новую переменную x_3: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$.

Для ответа на вопрос является ли переменная x_3 существенной необходимо сравнить следующие характеристики моделей:

а) Скорректированный коэффициент детерминации R^2_{adj}.

Коэффициент детерминации R^2 показывает долю дисперсии объясненную регрессией. Чем больше факторов включается в модель, тем большая часть дисперсии объясняется, поэтому при включении даже несущественной переменной можно заметить увеличение коэффициента детерминации R^2. Для устранения этого эффекта при анализе переменных используют скорректированный коэффициент детерминации:

$R^2_{adj} = 1 - (1 - R^2)\dfrac{n-1}{n-k-1}$, где n – количество наблюдений, k – количество объясняющих переменных.

R^2_{adj} корректируется в сторону уменьшения с ростом числа объясняющих переменных и увеличивается при добавлении новой переменной только тогда, когда добавленная переменная является значимой.

б) Статистика Фишера (F – статистика).

F – статистика позволяет проверить гипотезу об обоснованности включения новых переменных в модель.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_1 : \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0$$

Нулевая гипотеза принимается в том случае если расчетное значение F – статистики меньше соответствующего критического значения статистики Фишера, найденного по таблице критических точек распределения. Чем больше наблюдаемое значение F - статистики, тем более значима вся совокупность параметров регрессии.

Итак, если при включении переменной x_3 F – статистика увеличивается, то эту переменную можно считать существенной.

в) Среднее квадратическое отклонение σ.

Среднее квадратическое отклонение показывает среднюю величину отклонения реальных значений от линии регрессии. Чем меньше отличаются эмпирические и теоретические значения, тем меньше значение σ и качественнее модель.

С добавлением значимой переменной в модель значение σ уменьшается.

Итак, если переменная x_3 значима, то в первой модели имеем неверную спецификацию. Оценки параметров β_1, β_2 будут смещены от истинного значения.

Если в уравнении регрессии имеется несущественная переменная, то она обнаружит себя по низкой t – статистике. В дальнейшем эту переменную следует исключить из рассмотрения. Однако добавлять и исключать переменные целесообразно по одной и сравнивать качество полученных уравнений регрессии. При сравнении качества двух моделей обязательным является требование, чтобы зависимая переменная была представлена в одной и той же форме и число наблюдений для обеих моделей было одинаковым.

Пример

Определить вид уравнения регрессии, описывающей зависимость между переменными оценить параметры регрессии, ее качество. Данные представлены в таблице 2

Таблица 2

x - спрос	5	7	10	12	15	20	21	25	26	30
y - доход	3,51	9,5	27,4	47	91	213	245	412	462	705

Решение Построим корреляционное поле, изобразив значения x и y в системе координат (рис. .

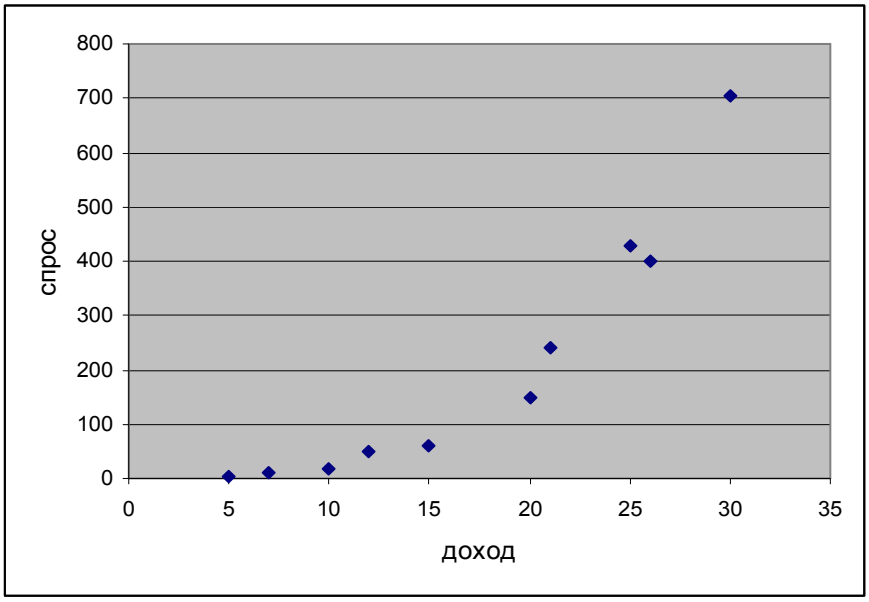

Рис. 5

По корреляционному полю замечаем, что зависимость может быть описана функциями: $y = a \cdot e^{bx}$; $y = ax^b$; $y = ax^2 + bx + c$ и др.

Однако, по условию задачи переменная y – это спрос, а x – доход. Из экономической теории известно, что зависимость между спросом на некоторое благо и доходом описывается функцией Энгеля вида $y = ax^b$. Поэтому уравнение регрессии для данных переменных y и x, будем искать в виде $y = ax^b$.

Линеаризируем функцию путем математических преобразований (логарифмирования).

$$y = ax^b$$

$$\ln y = \ln ax^b$$

$$\ln y = \ln a + \ln x^b$$

$$\ln y = \ln a + b \ln x$$

Обозначим $Y = \ln y$, $X = \ln x$, $A = \ln a$, получим

$$Y = A + bX$$

Оценим параметры полученного линейного уравнения методом наименьших квадратов. Необходимые расчёты представим в таблице 3.

Таблица 3

№	x	y	X=ln x	Y=ln y	X^2	XY
1	5	3,51	1,61	1,26	2,59	2,02
2	7	9,5	1,95	2,25	3,79	4,38
3	10	17,8	2,30	2,88	5,30	6,63
4	12	50	2,48	3,91	6,17	9,72
5	15	60	2,71	4,09	7,33	11,09
6	20	150	3,00	5,01	8,97	15,01
7	21	240	3,04	5,48	9,27	16,69
8	25	430	3,22	6,06	10,36	19,52
9	26	400	3,26	5,99	10,62	19,52
10	30	705	3,40	6,56	11,57	22,31
сумма	171	2065,81	26,97	43,50	75,98	126,88
среднее	17,1	206,58	2,70	4,35	7,60	12,69

Для нахождения оценок параметров *A* и *b* составим систему уравнений:

$$\begin{cases} A + b \cdot \overline{X} = \overline{Y} \\ A \cdot \overline{X} + b \cdot \overline{X^2} = \overline{X \cdot Y} \end{cases} \qquad \begin{cases} A + 2{,}7 \cdot b = 4{,}35 \\ 2{,}7 \cdot A + 7{,}6 \cdot b = 12{,}69 \end{cases}$$

Решая систему уравнений получаем: *A=-3,61; b=2,95*.

Зная, что $A = \ln a$ определим оценку параметра $a = e^A = e^{-3{,}61} = 0{,}027$, тогда уравнение зависимости примет вид: $y = 0{,}027 \cdot x^{2{,}95}$

Оценим качество полученной регрессии с помощью коэффициента детерминации: $R^2 = 1 - \frac{\sum (y_i - y_i)^2}{\sum (y_i - \overline{y})^2}$. Промежуточные вычисления представим в таблице 4.

Таблица 4

	x	y	$y = 0{,}027 \cdot x^{2{,}95}$	$(y - y)^2$	$(y - \overline{y})^2$
1	5	3,51	3,11	0,16	41237,83
2	7	9,5	8,40	1,20	38840,92
3	10	17,8	24,06	39,23	35638,27
4	12	50	41,20	77,35	24517,61
5	15	60	79,59	383,58	21485,99
6	20	150	185,95	1292,59	3201,41
7	21	240	214,74	638,12	1116,83
8	25	430	359,16	5018,44	49916,05
9	26	400	403,21	10,33	37410,91
10	30	705	614,99	8100,93	248421,50
сумма		2065,81		15561,95	501787,3
среднее		206,58			

$$R^2 = 1 - \frac{15561{,}95}{501787{,}3} = 0{,}967$$

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k} = \frac{0{,}967}{1 - 0.967} \cdot \frac{10 - 1 - 1}{1} = 249{,}96$$

Значение коэффициента детерминации близко к 1, F-статистика указывает на значимость коэффициента детерминации. Полученная модель имеет отличное качество.

Коэффициент *b=2,95* в данной модели является коэффициентом эластичности, т.е. показывает, что с увеличением дохода потребителя на 1% спрос на

товар увеличивается на *2,95* %.

Задание для самостоятельной работы

Имеются данные о стоимости однокомнатных квартир (таблица 5), где Price - цена, тыс. руб.; So - общая площадь, $м^2$; R - расстояние до центра, км.

Таблица 5

№	PRICE	SO	R
1	460	46	7
2	350	44	3,5
3	490	57,6	5
4	470	53,1	1,2
5	350	50	5
6	450	52,1	1,5
7	300	48	9
8	370	53	6,5
9	380	49	5
10	430	42	1,2
11	400	44	2
12	420	41	0

Оцените регрессии: $Price=\alpha+\beta*So$; $Price=\alpha+\beta*R$, $Price=\alpha+\beta*R+\gamma*So$ дайте объяснение полученным результатам.

3 КЛАСТЕРНЫЙ АНАЛИЗ

3.1 Постановка задачи

Классификация является основой умозрительной человеческой деятельности и фундаментальным процессом научной практики. В ходе исследований, развития науки и техники накоплено значительное количество материалов, которые необходимо систематизировать с целью выявления законов общественного развития, изучения эволюций и совершенствования технологий. Эта работа требует от исследователя детального изучения данных и их обобщения, в ходе которого отдельные факты складываются в закономерности, а закономерности в теории. Общий вопрос, задаваемый исследователями во многих областях, состоит в том, как организовать наблюдаемые данные в наглядные структуры. В настоящее время существует множество подходов к классификации объектов.

Среди них кластерный анализ - наиболее действенный количественный

инструмент исследования социально - экономических процессов, описываемых большим числом характеристик.

Кластерный анализ наиболее ярко отражает черты многомерного анализа в классификации. Термин кластерный анализ включает в себя набор различных алгоритмов классификации.

Первое применение кластерный анализ нашел в социологии. Название кластерный анализ происходит от английского слова cluster - гроздь, скопление. Впервые в 1939 был определен предмет кластерного анализа и сделано его описание исследователем Трионом. Главное назначение кластерного анализа - разбиение множества исследуемых объектов и признаков на однородные в соответствующем понимании группы или кластеры. Это означает, что решается задача классификации данных и выявления соответствующей структуры в ней.

Методы кластерного анализа позволяют решать следующие задачи:

1. Проведение классификации объектов с учетом признаков, отражающих сущность, природу объектов. Решение такой задачи, как правило, приводит к углублению знаний о совокупности классифицируемых объектов;

2. Проверка выдвигаемых предположений о наличии некоторой структуры в изучаемой совокупности объектов, т. е. поиск существующей структуры;

3. Построение новых классификаций для слабоизученных явлений, когда необходимо установить наличие связей внутри совокупности и попытаться привнести в нее структуру.

Большое достоинство кластерного анализа в том, что он позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов, и позволяет рассматривать множество исходных данных практически произвольной природы. Это имеет большое значение, например, для прогнозирования конъюнктуры, когда показатели имеют разнообразный вид, затрудняющий применение традиционных эконометрических подходов.

Кластерный анализ позволяет рассматривать достаточно большой объем

информации и резко сокращать, сжимать большие массивы информации, делать их компактными и наглядными.

Важное значение кластерный анализ имеет применительно к совокупностям временных рядов, характеризующих экономическое развитие. Здесь можно выделять периоды, когда значения соответствующих показателей были достаточно близкими, а также определять группы временных рядов, динамика которых наиболее схожа.

Кластерный анализ можно использовать циклически. В этом случае исследование производится до тех пор, пока не будут достигнуты необходимые результаты. При этом каждый цикл может давать информацию, которая способна сильно изменить направленность и подходы дальнейшего применения кластерного анализа. Этот процесс можно представить системой с обратной связью.

В задачах социально-экономического прогнозирования весьма перспективно сочетание кластерного анализа с другими количественными методами (например, с регрессионным анализом).

Как и любой другой метод, кластерный анализ имеет определенные недостатки и ограничения. В частности, состав и количество кластеров зависит от выбираемых критериев разбиения. При сведении исходного массива данных к более компактному виду могут возникать определенные искажения, а также могут теряться индивидуальные черты отдельных объектов за счет замены их характеристиками обобщенных значений параметров кластера. При проведении классификации объектов игнорируется очень часто возможность отсутствия в рассматриваемой совокупности каких-либо значений кластеров.

Поэтому необходимо сделать несколько предостережений общего характера.

1) Многие методы кластерного анализа - довольно простые эвристические процедуры, которые, как правило, не имеют достаточного статистического обоснования.

2) Разные кластерные методы могут порождать и порождают различные

решения для одних и тех же данных. Это обычное явление в большинстве прикладных исследований.

3) Цель кластерного анализа заключается в поиске существующих структур. В то же время его действие состоит в привнесении структуры в анализируемые данные, т. е. методы кластеризации могут приводить к порождению артефактов.

Исследования, использующие кластерный анализ, характеризуют следующие пять основных шагов: 1) отбор выборки для кластеризации; 2) определение множества признаков, по которым будут оцениваться объекты в выборке, и способа их стандартизации; 3) вычисление значений той или иной меры сходства между объектами; 4) применение метода кластерного анализа для создания групп сходных объектов; 5) проверка достоверности результатов кластерного решения.

В кластерном анализе считается, что:

а) выбранные характеристики допускают в принципе желательное разбиение на кластеры;

б) единицы измерения (масштаб) выбраны правильно.

Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве *X,* разбить множество объектов *G* на *m* (*m* - целое) кластеров (подмножеств) $Q_1, Q_2, ..., Q_m$, так, чтобы каждый объект G_j принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие одному и тому же кластеру, были сходными, в то время, как объекты, принадлежащие разным кластерам были разнородными.

3.2 Иерархические кластерные структуры

Решением задачи кластерного анализа являются разбиения, удовлетворяющие некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и группировок, который называют целевой функцией. Например, в качестве целевой функции может быть взята внутригрупповая сумма

квадратов отклонений:

$$W = \sum_{j=1}^{n}(x_j - \bar{x})^2 = \sum_{j=1}^{n} x_j^2 - \frac{1}{n}(\sum_{j=1}^{n} x_j)^2$$

где x_j - представляет собой измерения j - го объекта.

Наиболее трудным в задаче классификации является определение меры однородности объектов.

Понятно, что объекты *i*-ый и *j*-ый попадали бы в один кластер, когда расстояние (отдаленность) между точками X_i и X_j было бы достаточно маленьким и попадали бы в разные кластеры, когда это расстояние было бы достаточно большим. Таким образом, попадание в один или разные кластеры объектов определяется понятием расстояния между X_i и X_j из E_p, где E_p p - мерное евклидово пространство.

Неотрицательная функция $\rho(X_i, X_j)$ называется функцией расстояния (метрикой), если:

а) $\rho(X_i, X_j) \geq 0$, для всех X_i и X_j из E_p

б) $\rho(X_i, X_j) = 0$, тогда и только тогда, когда $X_i = X_j$

в) $\rho(X_i, X_j) = \rho(X_j, X_i)$

г) $\rho(X_i, X_j) \leq \rho(X_i, X_k) + \rho(X_k, X_j)$, где X_i; X_j и X_k - любые три вектора из E_p.

Значение $\rho(X_i, X_j)$ для X_i и X_j называется расстоянием между X_i и X_j и эквивалентно расстоянию между G_i и G_j соответственно выбранным характеристикам $(F_1, F_2, F_3, ..., F_p)$.

Наиболее часто употребляются следующие функции расстояний:

$\rho = \sqrt{\sum (x-y)^2}$ – евклидово расстояние, наиболее общий тип расстояния. Оно является геометрическим расстоянием в многомерном пространстве;

$\rho = \sum (x-y)^2$ – квадрат евклидова расстояния используется для того, чтобы

придать большие веса более отдаленным друг от друга объектам;

$\rho = \sum |x - y|$ – расстояние городских кварталов (манхэттенское расстояние) для этой меры влияние отдельных больших разностей (выбросов) уменьшается;

$\rho = \max(x - y)$ – расстояние Чебышева полезно, когда желают определить два объекта как «различные», если они различаются по какой-либо одной координате (каким-либо одним измерением);

$\rho = \left(\sum |x - y|^p\right)^{\frac{1}{k}}$ степенное расстояние используют, когда хотят увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Параметры r и p определяются пользователем Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, параметр r ответственен за прогрессивное взвешивание больших расстояний между объектами;

$\rho = \frac{\textit{количество } x_i \neq y_i}{i}$ – процент несогласия используется в тех случаях, когда данные являются категориальными.

Пусть n измерений $X_1, X_2, ..., X_n$ представлены в виде матрицы данных размером $p \times n$

$$X = \begin{pmatrix} x_{11} & x_{12} & ..x_{1n} \\ x_{21} & x_{22} & ..x_{2n} \\ x_{p1} & x_{p2} & ..x_{pn} \end{pmatrix} = \left(X_1, X_2, ..., X_n\right)$$

Тогда расстояние между парами векторов $\rho\left(X_i, X_j\right)$ могут быть представлены в виде симметричной матрицы расстояний:

$$D = \begin{pmatrix} 0 & \rho_{12} & ... & \rho_{1n} \\ \rho_{21} & 0 & ... & \rho_n \\ \rho_{n1} & \rho_{n2} & ... & 0 \end{pmatrix}$$

Понятием, противоположным расстоянию, является понятие сходства между объектами G_l и G_j. Неотрицательная вещественная функция

$S(X_i; X_j) = s_{ij}$ называется мерой сходства, если:

1) $0 \le S(X_i, X_j) < 1$ для $X_l \ne X_j$

2) $S(X_i, X_i) = 1$

3) $S(X_i, X_j) = S(X_j, X_i)$

Пары значений мер сходства можно объединить в матрицу сходства:

$$S = \begin{pmatrix} 1 & s_{12} & \dots & s_{1n} \\ s_{21} & 1 & \dots & s_{2n} \\ s_{n1} & s_{n2} & \dots & 1 \end{pmatrix}$$

Величину s_{ij} называют коэффициентом сходства.

Естественной мерой сходства характеристик объектов во многих задачах является коэффициент корреляции между ними

$$r_{ij} = \frac{\sum_{h=1}^{N} (x_{hi} - m_i)(x_{hj} - m_j)}{\sigma_i \cdot \sigma_j}$$, где $m_i, m_j, \sigma_i, \sigma_j$ - соответственно средние и среднеквадратичные отклонения для характеристик *i* и *j*. Мерой различия между характеристиками может служить величина *1-r*.

На первом шаге, когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Однако когда связываются вместе несколько объектов необходимо правило объединения или связи для двух кластеров. Существует множество методов объединения кластеров, перечислим наиболее распространенные:

Одиночная связь (метод ближайшего соседа) – расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами.

Полная связь (метод наиболее удаленных соседей) – расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах.

Невзвешенное попарное среднее – расстояние между двумя различными

кластерами вычисляется как среднее расстояние между всеми парами объектов в них.

Взвешенное попарное среднее – идентично методу невзвешенного попарного среднего, за исключением того, что при вычислениях размер соответствующих кластеров (т.е. число объектов, содержащихся в них) используется в качестве весового коэффициента.

Невзвешенный центроидный метод – расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.

Взвешенный центроидный метод (медиана) – идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. числами объектов в них)

Метод Варда. – отличается от всех других методов, поскольку он использует методы дисперсионного анализа для оценки расстояний между кластерами. Метод минимизирует сумму квадратов (SS) для любых двух кластеров, которые могут быть сформированы на каждом шаге. В целом метод представляется очень эффективным, однако он стремится создавать кластеры малого размера.

Число алгоритмов методов кластерного анализа слишком велико. Все их можно подразделить на иерархические и неиерархические.

Иерархические алгоритмы связаны с построением дендограмм и делятся на:

а) агломеративные, характеризуемые последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров;

б) дивизимные (делимые), в которых число кластеров возрастает, начиная с одного, в результате чего образуется последовательность расщепляющих групп.

Иерархические агломеративные методы – многошаговые методы, работающие в такой последовательности: на нулевом шаге за разбиение принимается исходная совокупность n элементарных кластеров, матрица расстояний между которыми $\{\rho_{ij}\}_{n\times n} = \{l_{ij}\}_{n\times n}$; на каждом следующем шаге происходит объеди-

нение двух кластеров K_s и K_t, сформированных на предыдущем шаге, в один кластер $K_s \cup K_t$ (будем его обозначать $K_{s \oplus t}$, при этом размерность матрицы расстояний уменьшается, по сравнению с размерностью матрицы предыдущего шага, на единицу.

Наиболее известный метод представления матрицы расстояний или сходства основан на идее дендрограммы или диаграммы дерева. Дендрограмму можно определить как графическое изображение результатов процесса последовательной кластеризации, которая осуществляется в терминах матрицы расстояний. С помощью дендрограммы можно графически или геометрически изобразить процедуру кластеризации при условии, что эта процедура оперирует только с элементами матрицы расстояний или сходства.

Существует много способов построения дендрограмм. В дендрограмме объекты располагаются вертикально слева, результаты кластеризации - справа. Значения расстояний или сходства, отвечающие строению новых кластеров, изображаются по горизонтальной прямой поверх дендрограмм.

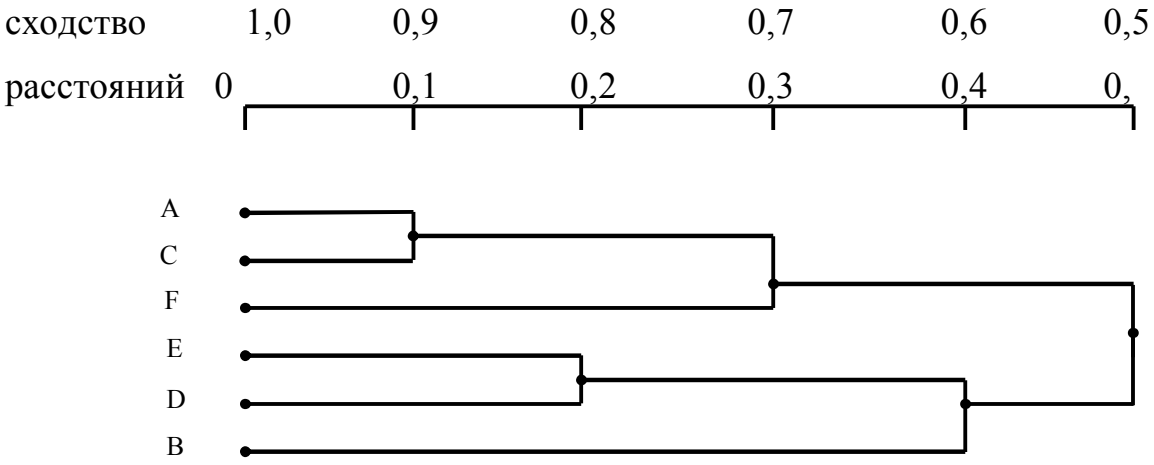

Рис. 6

На рисунке 6 показан один из примеров дендрограммы. Он соответствует случаю шести объектов *(n=6)* и *k* характеристик (признаков). Объекты *A* и *C* наиболее близки и поэтому объединяются в один кластер на уровне близости,

равном 0,9. Объекты D и E объединяются при уровне 0.8. Теперь имеем 4 кластера:

$$(A \oplus C), (F), (D \oplus E), (B).$$

Далее образуются кластеры *((A⊕C)⊕F)* и *((E⊕D), B),* соответствующие уровню близости, равному 0,7 и 0,6. Окончательно все объекты группируются в один кластер при уровне 0,5.

Вид дендрограммы зависит от выбора меры сходства или расстояния между объектом и кластером и метода кластеризации.

3.3 Метод k-средних

Иерархические методы используются обычно в задачах классификации небольшого числа объектов (порядка нескольких десятков), где больший интерес представляет не число кластеров, а анализ структуры множества этих объектов и наглядная интерпретация проведенного анализа в виде дендрограммы. Если же число кластеров заранее задано или подлежит определению, то для классификации чаще всего используют *параллельные* кластер-процедуры – это итерационные алгоритмы, на каждом шаге которых *используется* одновременно (параллельно) все наблюдения. Так как эти алгоритмы на каждом шаге работают со всеми наблюдениями, то основной целью их конструирования является нахождение способов сокращения числа перебора вариантов (даже при числе наблюдений порядка нескольких десятков), что приводит зачастую лишь к приближенному, но не слишком трудоемкому решению задач. В параллельных кластер-процедурах реализуется обычно идея оптимизации разбиения в соответствии с некоторым функционалом качества.

Наиболее распространенными являются при заданном числе k кластеров следующие функционалы качества разбиения:

- сумма внутрикластерных дисперсий $f(R) = \sum_{p=1}^{k} \sum_{i \in K_p} l^2\left(x_i, \bar{x}_{K_p}\right),$

– сумма попарных внутрикластерных расстояний $f(R)=\sum_{p=1}^{k}\frac{1}{n_p}\sum_{i,j\in K_p} l^2(x_i,x_j)$, а при неизвестном числе кластеров функционалы

$f(R)=\alpha f_1(R)+\beta f_2(R)$ или $f(R)=[f_1(R)]^{\alpha}[f_2(R)]^{\beta}$,

где $f_1(R)$ - некоторая не возрастающая функция числа классов, характеризующая средний внутриклассовый разброс наблюдений, $f_2(R)$ - некоторая неубывающая функция числа классов, характеризующая взаимную удаленность классов или меру «концентрации» наблюдений.

Схема работы алгоритмов, связанная с функционалами качества, такая: для некоторого начального разбиения R_0 вычисляют значение $f(R_0)$; затем каждую из точек x_i, поочередно перемещают во все кластеры и оставляют в том положении, которое соответствует наилучшему значению функционала качества. Работу заканчивают, когда перемещение точек не дает улучшения качества. Часто описанный алгоритм применяют несколько раз, начиная с разных начальных разбиений R_0, и выбирают наилучший вариант разбиения.

Очень важным вопросом является проблема выбора необходимого числа кластеров. Иногда можно число кластеров выбирать априорно. Однако в общем случае это число определяется в процессе разбиения множества на кластеры.

Проводились исследования Фортьером и Соломоном, и было установлено, что число кластеров должно быть принято для достижения вероятности a того, что найдено наилучшее разбиение. Таким образом, оптимальное число разбиений является функцией заданной доли β наилучших или в некотором смысле допустимых разбиений во множестве всех возможных. Общее рассеяние будет тем больше, чем выше доля β допустимых разбиений. Фортьер и Соломон разработали таблицу, по которой можно найти число необходимых разбиений $S(\alpha,\beta)$ в зависимости от α и β (где α - вероятность того, что найдено наилучшее разбиение, β - доля наилучших разбиений в общем числе разбиений) Причем в качестве меры разнородности используется не мера

рассеяния, а мера принадлежности, введенная Хользенгером и Харманом. Значения $S(\alpha,\beta)$ приводится в таблице 6.

Довольно часто критерием объединения (числа кластеров) становится изменение соответствующей функции. Например, суммы квадратов отклонений:

$$E_j = \sum_{i=1}^{n} r_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} r_{ij}\right)^2 .$$

Таблица 6

$\beta \setminus \alpha$	0.20	0.10	0.05	0.01	0.001	0.0001
0.20	8	11	14	21	31	42
0.10	16	22	29	44	66	88
0.05	32	45	59	90	135	180
0.01	161	230	299	459	689	918
0.001	1626	2326	3026	4652	6977	9303
0.0001	17475	25000	32526	55000	75000	100000

Процессу группировки должно соответствовать здесь последовательное минимальное возрастание значения критерия *E*. Наличие резкого скачка в значении *E* можно интерпретировать как характеристику числа кластеров, объективно существующих в исследуемой совокупности.

Итак, второй способ определения наилучшего числа кластеров сводится к выявлению скачков, определяемых фазовым переходом от сильно связанного к слабосвязанному состоянию объектов.

Иерархические и параллельные кластер-процедуры практически реализуемы лишь в задачах классификации не более нескольких десятков наблюдений. К решению задач с большим числом наблюдений применяют *последовательные* кластер-процедуры - это итерационные алгоритмы, на каждом шаге которых используется одно наблюдение (или небольшая часть исходных наблюдений) и результаты разбиения на предыдущем шаге. Идею этих процедур поясним на представленном в ППП «STASTICA» *методе К-средних* («*К* – Means Clusyering») с заранее заданным числом *k* классов.

На нулевом шаге за центры искомых *k* кластеров принимают случайно выбранные *k* наблюдений – точки $x_1, x_2, \ldots x_k$; каждому кластеру присваивают

единичный вес. На первом шаге находят расстояния точки x_{k+1} до центров кластеров и точку x_{k+1} относят к кластеру, расстояние до которого минимально; рассчитывают новый центр тяжести (как взвешенное среднее по каждому показателю) этого кластера и вес кластера увеличивают на единицу; все остальные кластеры остаются неизмененными (с прежними центрами и весами). На втором шаге аналогичную процедуру выполняют для точки x_{k+2} и т. д. При достаточно большом числе n классифицируемых объектов или достаточно большом числе итерации пересчет центров тяжести практически не приводит к их изменению.

Если в какой-то точке не удается, прогнав все $x_{k+(n-1)}$ точек, достичь практически не изменяющихся центров тяжести, то либо используя получившееся разбиение n точек на k кластеров в качестве начального применяют изложенную процедуру к точкам x_1, x_2 и т. д.; либо в качестве начального разбиения принимают различные комбинации k точек из исходных n точек и в качестве окончательного берут наиболее часто встречающееся финальное разбиение.

Кластерный анализ методом к-средних дополняет и уточняет картину, полученную с помощью иерархического кластерного анализа. Однако конфигурация кластеров не поддается представлению в графическом виде.

Примеры

1. Провести классификации пяти точек: (1,2), (4, 3), (-1, -1), (-1, 0), (-3, 3):

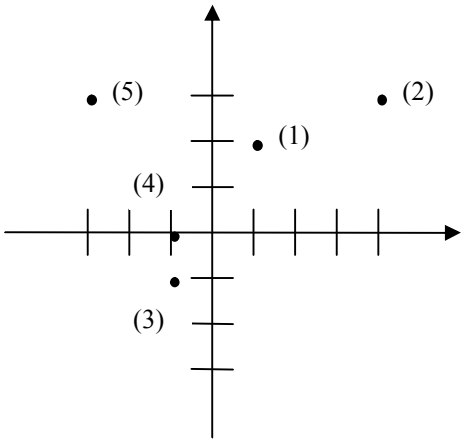

Рис. 7

Решение

За метрику расстояний примем квадратичное евклидово расстояние.

Например, между точками (1) и (2) евклидово расстояние составит: $\rho_{1,2} = (1-4)^2 + (2-3)^2 = 10$.

Тогда матрица расстояний между точками примет вид (табл. 7):

Таблица 7

	1	2	3	4	5
1	0	10	13	8	17
2	10	0	41	34	49
3	13	41	0	1	20
4	8	34	1	0	13
5	17	49	20	13	0

Начальное разбиение: 1, 2, 3, 4, 5. Дальнейшее разбиение проведем методом ближайшего соседа. Минимальное расстояние $\rho_{3,4} = 1$, переходим к разбиению: 1, 2, $3 \oplus 4$, 5.

Последовательность разбиений: 1, 2, 3, 4, $5 \to 1$, 2, $3 \oplus 4$, $5 \to 1 \oplus (3 \oplus 4)$, 2, $5 \to (1 \oplus (3 \oplus 4)) \oplus 2$, $5 \to 1 \oplus 2 \oplus 3 \oplus 4 \oplus 5$. Ее результаты показаны на схеме:

	1	2	$3 \oplus 4$	5
1	0	10	8	17
2	10	0	34	49
$3 \oplus 4$	8	34	0	13
5	17	49	13	0

$\to$

	$1 \oplus (3 \oplus 4)$	2	5
$1 \oplus (3 \oplus 4)$	0	10	13
2	10	0	49
5	13	49	0

$\to$

	$(1 \oplus (3 \oplus 4)) \oplus 2$	5
$(1 \oplus (3 \oplus 4)) \oplus 2$	0	13
5	13	0

Поскольку этот метод объединяет кластеры, в которых расстояние между ближайшими элементами минимально по сравнению с другими кластерами, то два объекта попадают в один кластер, если существует соединяющая их цепочка ближайших друг к другу объектов («цепочечный эффект»). Поэтому метод «ближний сосед» называют методом «одиночной связи». Для устранения «цепочечного эффекта» можно, например, ввести ограничение на максимальное расстояние между объектами кластера: в первый кластер включить два наиболее близких объекта, затем в этот кластер включить объект, который имеет минимальное расстояние с одним из объектов кластера, а его расстояние до друго-

го объекта кластера не больше числа l_0 и т. д.; формирование первого кластера продолжают до тех пор, пока нельзя будет найти объект, расстояние которого до любого объекта кластера, не превзойдет l_0; формирование 2-го и последующих кластеров осуществляется из оставшихся объектов аналогичным образом.

Рассмотрим процедуру кластерного анализа другими методами.

а) Дальний сосед

	1	2	$3\oplus4$	5
1	0	10	13	17
2	10	0	41	49
$3\oplus4$	13	41	0	20
5	17	49	20	0

$\rightarrow$

	$1\oplus2$	$3\oplus4$	5
$1\oplus2$	0	41	49
$3\oplus4$	41	0	20
5	49	20	0

$\rightarrow$

	$1\oplus2$	$(3\oplus4)\oplus5$
$1\oplus2$	0	49
$(3\oplus4)\oplus5$	49	0

Последовательность разбиений: 1, 2, 3, 4, 5 $\rightarrow 1,\ 2,\ 3\oplus4,\ 5 \rightarrow 1\oplus2,\ 3\oplus4,\ 5 \rightarrow 1\oplus2,\ (3\oplus4)\oplus5 \rightarrow 1\oplus2\oplus3\oplus4\oplus5$.

В этом методе объединяются кластеры, в которых минимально расстояние между самыми далекими друг от друга объектами. Это означает, что все остальные объекты в полученном после объединения кластере связаны друг с другом еще теснее, чем «соседи». Поэтому метод «дальнего соседа» называют методом полной связи.

б) Средняя связь

	1	2	$3\oplus4$	5
1	0	10	10.5	17
2	10	0	37.5	49
$3\oplus4$	10.5	37.5	0	16.5
5	17	49	16.5	0

$\rightarrow$

	$1\oplus2$	$3\oplus4$	5
$1\oplus2$	0	24	33
$3\oplus4$	24	0	16.5
5	33	16.5	0

$\rightarrow$

	$1\oplus2$	$(3\oplus4)\oplus5$
$1\oplus2$	0	27
$(3\oplus4)\oplus5$	27	0

В последней таблице

$$\rho_{1\oplus2,(3\oplus4)\oplus5} = \frac{2}{3}\rho_{1\oplus2,3\oplus4} + \frac{1}{3}\rho_{1\oplus2,5} = \frac{2}{3}\cdot 24 + \frac{1}{3}\cdot 33 = 27$$

Последовательность разбиений: 1, 2, 3, 4, 5 $\rightarrow 1,\ 2,\ 3\oplus4,\ 5 \rightarrow 1\oplus2,$ $3\oplus4,\ 5 \rightarrow 1\oplus2, (3\oplus4)\oplus5 \rightarrow 1\oplus2\oplus3\oplus4\oplus5$.

в) Центроидный метод

	1	2	$3\oplus4$	5
1	0	10	10.25	17
2	10	0	37.25	49

$\rightarrow$

	$1\oplus2$	$3\oplus4$	5
$1\oplus2$	0	21.25	30.5
$3\oplus4$	21.25	0	16.25

$\rightarrow$

	$1\oplus2$	$(3\oplus4)\oplus5$
$1\oplus2$ $(3\oplus4)\oplus5$	0	$\frac{373}{18}$

$3 \oplus 4$	10.25	37.25	0	16.25
5	17	49	16.25	0

5	30.5	16.25	0

	$\frac{373}{18}$	0
	$1 \oplus 2$	$(3 \oplus 4) \oplus 5$

2. Построить дендрограмму для классификации стран по уровню экономического развития (таблица 8), если известны доли импорта и экспорта данных стран, общее количество населения, а также процент безработного населения каждой страны.

Таблица 8

страна	доля экспорта (%)	доля импорта (%)	население (млн.чел.)	% безработного населения
Россия	50	80	146	20
Украина	20	50	120	35
Белоруссия	30	25	100	15
Польша	10	60	95	10
Бельгия	60	20	70	15

Решение Данные необходимо стандартизировать, т.к. они имеют разные единицы измерения (табл. 9).

Таблица 9

страна	доля экспорта (%)	доля импорта (%)	население (млн.чел.)	% безработного населения
Россия	0,77	1,33	1,40	0,10
Украина	-0,68	0,12	0,48	1,66
Белоруссия	-0,19	-0,88	-0,22	-0,42
Польша	-1,16	0,52	-0,39	-0,94
Бельгия	1,25	-1,08	-1,27	-0,42

1 шаг. Все объекты располагаем на графике по оси OY, на оси OX расположим расстояния. За метрику расстояний примем евклидово расстояние. Матрица расстояний представлена в таблице 10.

Таблица 10

	Россия	Украина	Белоруссия	Польша	Бельгия
Россия	0	2,61	2,95	2,94	3,66
Украина	2,61	0	2,46	2,81	3,55

Белоруссия	2,95	2,46	0	1,79	1,80
Польша	2,94	2,81	1,79	0	3,07
Бельгия	3,66	3,55	1,80	3,07	0

Из данных расстояний выбираем наименьшее 1,79 и объединяем объекты в один класс. Получается 4 объекта: Россия, Украина, (Польша+Белоруссия), Бельгия.

2 шаг. В качестве меры связи выберем центроидный метод. Найдем значения для кластера (Польша+Белоруссия) как среднее арифметическое (табл. 11).

Таблица 11

страна	доля экспорта (%)	доля импорта (%)	население (млн.чел.)	% безработного населения
Россия	0,77	1,33	1,40	0,10
Украина	-0,68	0,12	0,48	1,66
Белоруссия+Польша	-0,68	-0,18	-0,31	-0,68
Бельгия	1,25	-1,08	-1,27	-0,42

Вновь рассчитаем расстояния (табл. 12).

Таблица 12

	Россия	Украина	Белоруссия+Польша	Бельгия
Россия	0,00	2,61	2,81	3,66
Украина	2,61	0,00	2,49	3,54
Белоруссия+Польша	2,81	2,49	0,00	2,35
Бельгия	3,66	3,55	2,35	0,00

Самое минимальное расстояние 2,35 между кластером (Белоруссия+Польша) и Бельгией. Объединяем их в одну группу и отображаем на графике. Процесс продолжаем до тех пор, пока все страны не объединятся в один кластер.

3 шаг.

Таблица 13

страна	доля экспорта (%)	доля импорта (%)	население (млн.чел.)	% безработного населения
Россия	0,77	1,33	1,40	0,10

Украина	-0,68	0,12	0,48	1,66
Белорус-сия+Польша+Бельгия	0,29	-0,63	-0,79	-0,55

Матрица расстояний представлена в таблице 14.

Таблица 14

	Россия	Украина	Белоруссия+Польша+Бельгия
Россия	0,00	2,61	3,05
Украина	2,61	0,00	2,83
Белоруссия+Польша+Бельгия	3,05	2,83	0,00

Минимальное расстояние 2,61, на котором объединяются Россия и Украина.

4 шаг.

Таблица 15

страна	доля экс-порта (%)	доля им-порта (%)	население (млн.чел.)	% безработного населения
Россия+Украина	0,05	0,73	0,94	0,88
Белорус-сия+Польша+Бельгия	0,29	-0,63	-0,79	-0,55

Матрица расстояний представлена в таблице 16.

Таблица 16.

	Россия+Украина	Белоруссия+Польша+Бельгия
Россия+Украина	0,00	3,05
Белоруссия+Польша+Бельгия	3,05	0,00

Итак все объекты объединяются на расстоянии 3,05.

Дендрограмма приведена на рисунке 8.

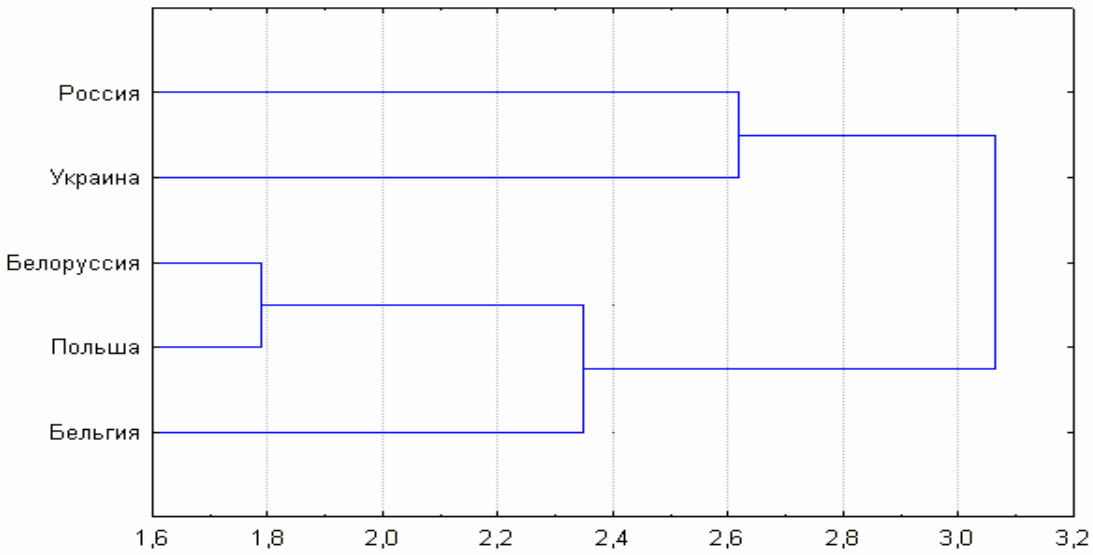

Рис. 8

Задание для самостоятельной работы

1. Используя не менее двух методов кластер – процедур провести классификацию точек (3; 4) , (-3; 8), (-2;-6), (5; 7), (6; 0), (8; 4), (-3; 9), (2;-6), (5; -7), (5; 0). В качестве расстояний выбрать евклидово расстояние и расстояние городских кварталов.

2. Объединить 8 Амурских фирм, занимающихся производством и установкой окон в несколько схожих совокупностей путем построения дендрограммы (табл. 17).

Таблица 17

Фирма	Возраст фирмы	Кол-во сотрудников	Количество установленных окон за месяц	Средняя цена окна (тыс.руб)	Количество филиалов по области	Скидки(%)	Популярность (0-нет, 1да)
Home master	11	200	30	17	20	25	1
Уют	3	14	12	13,2	2	33	1
РосОкна	3	50	8	14,8	5	30	1
Ремикс	5	80	14	15	10	10	1
Ванда	8	35	17	14,2	12	15	1
Ремстрой	2	10	5	11,5	1	5	0
СМУ-17	7	20	3	13,5	1	0	0
Оникс	3	20	1	12,9	1	0	0

4. ФАКТОРНЫЙ АНАЛИЗ, И ЕГО ИСПОЛЬЗОВАНИЕ В ИССЛЕДОВАНИИ СВЯЗИ

4.1 Постановка задачи

Факторный анализ – статистический метод, который используется при обработке больших массивов экспериментальных данных. Задачами факторного анализа являются: сокращение числа переменных (редукция данных) и определение структуры взаимосвязей между переменными, т.е. классификация переменных, поэтому факторный анализ используется как метод сокращения данных или как метод структурной классификации.

В современной статистике под факторным анализом понимают совокупность методов, которые на основе реально существующих связей признаков (или объектов) позволяют выявлять латентные обобщающие характеристики организационной структуры и механизма развития изучаемых явлений и процессов.

Понятие латентности в определении ключевое. Оно означает неявность характеристик, раскрываемых при помощи методов факторного анализа. Вначале мы имеем дело с набором элементарных признаков *Xj,* их взаимодействие предполагает наличие определенных причин, особенных условий, т.е. существование которых скрытых факторов. Последние устанавливаются в результате обобщения элементарных признаков и выступают как интегрированные характеристики, или признаки, но более высокого уровня. Естественно, что коррелировать могут не только тривиальные признаки X_j, но и сами наблюдаемые объекты N_i, Поэтому поиск латентных факторов теоретически возможен как по признаковым, так и по объектным данным.

Идея метода состоит в сжатии матрицы признаков в матрицу с меньшим числом переменных, сохраняющую почти ту же самую информацию, что и исходная матрица, т.е. сконцентрировать исходную информацию, выражая большое число рассматриваемых признаков через меньшее число более емких внутренних характеристик явления, которые, однако, не поддаются непосредственному измерению.

Предположим, n наблюдаемых объектов (автомобилей) оценивается в двумерном признаковом пространстве R^2 с координатными осями: X_1 – стоимость автомобиля и X_2 – длительность рабочего ресурса мотора. При условии коррелированности X_1 и X_2 в системе координат появляется направленное и достаточно плотное скопление точек, формально отображаемое новыми осями (F_1 и F_2). Характерная особенность F_1 и F_2 заключается в том, что они проходят через плотные скопления точек и в свою очередь коррелируют с X_1 и X_2. Максимальное число новых осей F_r будет равно числу элементарных признаков.

Допуская линейную зависимость F_r *от* X_{jt} можем записать:

$F_1 = a_{11} x_1 + a_{21} x_2$ и $F_2 = a_{12} x + a_{22} x_2$.

Интерпретируем оси пусть F_1 – экономичность автомобиля, F_2 – его надежность в эксплуатации. Суждение об F_1, и F_2 базируется на оценке структуры латентных факторов, т.е. оценке весов X_1 и X_2 в F_r .

Если объекты характеризуются достаточно большим числом элементарных признаков ($m > 3$), то логично и другое предположение – о существовании плотных скоплений точек (признаков) в пространстве n объектов. При этом новые оси обобщают уже не признаки X_{jt} а объекты, соответственно и латентные факторы F, будут распознаны по составу наблюдаемых объектов.

Материалом для факторного анализа служат корреляционные связи, а точнее – коэффициенты корреляции Пирсона, которые вычисляются между переменными, включенными в обследование.

В зависимости от того, какой тип корреляционной связи – элементарных признаков или наблюдаемых объектов – исследуется в факторном анализе, различают R и Q – технические приемы обработки данных.

Название R-техники носит объемный анализ данных по m признакам, в результате него получают r линейных комбинаций (групп) признаков ($F_r=f(Xj)$; r = l,n). Анализ по данным о близости (связи) n наблюдаемых объектов называется Q-техникой и позволяет определять r линейных комбинаций (групп) объектов.

В настоящее время на практике более 90% задач решается при помощи R-

техники.

Толчком для развития методов факторного анализа изначально послужили задачи и проблемы из области психологии. Позже методы факторного анализа стали активно применяться в социологических исследованиях, медицине, затем в военной промышленности, экономике.

Методы факторного анализа целесообразно разделить на два класса: упрощенные и современные аппроксимирующие методы.

Простые методы факторного анализа в основном связаны с начальными теоретическими разработками. Они имеют ограниченные возможности в выделении латентных факторов и аппроксимации факторных решений. В числе этих методов следует назвать:

однофакторную модель Ч. Спирмена. Она позволяет выделить только один генеральный латентный и один характерный факторы. Для возможно существующих других латентных факторов делается предположение об их незначимости;

бифакторную модель Г. Хользингера. Допускает влияние на вариацию элементарных признаков не одного, а нескольких латентных факторов (обычно двух) и одного характерного фактора;

центроидный метод Л. Тэрстоуна. В нем корреляции между переменными рассматриваются как пучок векторов, а латентный фактор геометрически представляется как уравновешивающий вектор, проходящий через центр этого пучка Метод позволяет выделять несколько латентных и характерные факторы, впервые появляется возможность соотносить факторное решение с исходными данными, т.е. простейшем виде решать задачу аппроксимации.

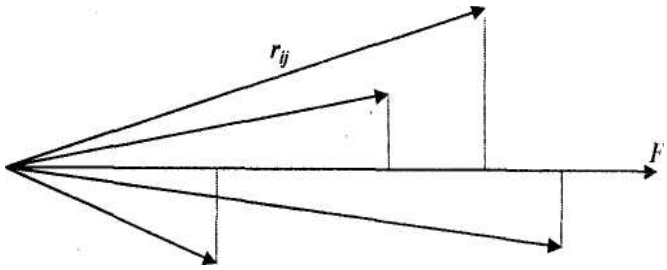

Рис. 9

Современные аппроксимирующие методы часто предполагают, что первое, приближенное решение уже найдено каким-либо из способов, последующими шагами это решение оптимизируется. Методы отличаются сложностью вычислений. К этим методам относятся:

• *групповой метод* Л. Гуттмана и П. Хорста. Решение базируется на предварительно отобранных каким-либо образомгруппах элементарных признаков;

• *метод главных факторов* Г. Томсона. Наиболее близок методу главных компонент, отличие заключается в предположении о существовании характерностей;

• *метод максимального правдоподобия* (Д. Лоули), *минимальных остатков* (Г. Харман), *а-факторного анализа* (Г. Кайзери И. Кэффри,), *канонического факторного анализа* (К. Рао),все *оптимизирующие.* Позволяют последовательно улучшить предварительно найденные решения на основе использования статистических приемов оценивания случайной величины или статистических критериев, предполагают большой объем трудоемких вычислений. Наиболее перспективным и удобным для работы в этой группе признается метод максимального правдоподобия.

Основной задачей, которую решают разнообразными методами факторного анализа, является сжатие информации, переход от множества значений по m элементарным признакам с объемом информации $n \times m$ к ограниченному множеству элементов матрицы факторного отображения $(m \times r)$ или матрицы значений латентных факторов для каждого наблюдаемого объекта размерностью $n \times r$, причем обычно $r < m$.

Методы факторного анализа позволяют также визуализировать структуру изучаемых явлений и процессов, а это значит определять их состояние и прогнозировать развитие. Наконец, данные факторного анализа дают основания для идентификации объекта, т.е. решения задачи распознавания образа.

Методы факторного анализа обладают свойствами, весьма привлекатель-

ными для их использования в составе других статистических методов, наиболее часто в корреляционно-регрессионном анализе, кластерном анализе, многомерном шкалировании и др.

Выделение латентных переменных, факторные нагрузки и факторные весы.

Главное понятие факторного анализа – фактор. Это искусственный статистический показатель, возникающий в результате специальных преобразований коэффициентов корреляции между изучаемыми признаками – латентная переменная.

Независимо от выбранного метода факторного анализа основные его результаты выражаются в наборах факторных нагрузок и факторных весов.

Факторные нагрузки - это значения коэффициентов корреляции каждого из исходных признаков с каждым из выявленных факторов. Чем теснее связь данного признака с рассматриваемым фактором, тем выше значение факторной нагрузки. Положительный знак факторной нагрузки указывает на прямую (а отрицательный знак - на обратную) связь данного признака с фактором. Таблица факторных нагрузок содержит *m* строк (по числу признаков) и k столбцов (по числу факторов).

Факторными весами называют количественные значения выделенных факторов для каждого из *n.* имеющихся объектов. Объекту с большим значением факторного веса присуща большая степень проявления свойств, определяемых данным фактором.

Поэтому положительные факторные веса соответствуют тем объектам, которые обладают степенью проявления свойств больше средней, а отрицательные факторные веса соответствуют тем объектам, для которых степень проявления свойств меньше средней. Таблица факторных весов содержит n строк (по числу объектов) и k столбцов (по числу факторов).

Таким образом, данные о факторных нагрузках позволяют сформулировать выводы о наборе исходных признаков, отражающих тот или иной фактор, и об относительном весе отдельного признака в структуре каждого фактора. В

свою очередь, данные о факторных весах определяют ранжировку объектов по каждому фактору.

В основе каждого метода факторного анализа лежит математическая модель, описывающая соотношения между исходными признаками и обобщенными латентными факторами.

Изучение факторных воздействий предполагает выявление взаимосвязей характерных признаков. Для многомерных объектов показателями связи являются оценки дисперсии и коэффициенты ковариации, которые обобщаются в матрице ковариаций (по выборочным данным – матрица *S).* Когда исходные значения признаков нормированы, матрица ковариаций, переходит в матрицу парных корреляций *R*.

$$S = R = \frac{1}{n} Z^T Z$$

Симметрическая матрица *R* имеет собственную систему координат в пространстве R^m, где m – число анализируемых признаков. Допуская преобразования координатной системы в систему пространства латентных факторов, можно записать Z в виде линейной комбинации новых координат в матричной форме: Z = *AF*.

Воспользуемся возможностью подстановки в уравнение для *R* вместо Z произведения матриц *AF* и получим:

$$R = \frac{1}{n} AF(AF)^T = \frac{1}{n} AFF^T A^T$$

Изменив место расположения скаляра 1/n, выделим произведение $\frac{1}{n} FF^T$, результат произведения интерпретируется как матрица корреляций C, определяемая для латентных факторов F_r. После замены 1/n *FF* на C запишем: $R = ACA^T$.

В предположении, что факторы F_r некоррелированы, т.е. $C = E$, где E – единичная матрица, приходим к равенству: $R = AA'$.

Л.Л. Тэрстоуном равенства типа: $R = ACA'$ и $R - AA'$ названы *фундаментальной факторной теоремой, A* – здесь матрица факторного отображения, а

ее элементы *a* – величины факторных нагрузок. Суть теоремы – в возможности воспроизведения исходной корреляционной матрицы *R* через матрицу факторного отображения *A*. При С = *E* связь матричных элементов г и *a* записывается в виде уравнения:

$$r_{ij} = a_{i1}a_{j1} + a_{i2}a_{j2} + \ldots + a_{ir}a_{jr}$$

Другими словами, корреляция пары характерных признаков опосредуется корреляцией каждого из признаков с некоторыми латентными факторами F_r. Латентные факторы определяют само существование связи i-го и j-го коррелирующих признаков.

Равенства Тэрстоуна допускаются гипотетически. Реально *AA'* и *ACA'* будут далеко не всегда в точности воспроизводить *R*. По крайней мере, это объясняется двумя причинами. Во-первых, в факторном анализе, позволяющем эффективно объяснять общую дисперсию данных, *r* – число латентных (обобщенных) признаков, как правило, значительно меньше числа исходных признаков *m*. И, во-вторых, в матрице *A* объединяются теоретические оценки факторных нагрузок. С учетом различий математических методов и специфичности вычислительных процедур следует допустить, что они не абсолютно истинны.

Таким образом, можно ожидать, что воспроизведенная из *AA'* или *ACA'* матрица корреляций R^+ будет отлична от R. Как следствие, на главной диагонали R^+ располагаются величины, обычно не равные, а меньшие единицы. На практике значения r^+_{ij} принимают за общности h_j, т.е. характеристики части дисперсии, поддавшейся объяснению через латентные факторы F_r, а $1 - a^+_{ij}$ – специфичность, т.е. необъясненная часть дисперсии. По степени расхождения R^+ и *R* судят о достаточности числа выделенных латентных факторов и адекватности аналитических выводов.

Матрица корреляций с общностями на главной диагонали называется редуцированной. Она является исходной для нахождения матрицы факторных нагрузок.

Существуют достаточно простые методы поиска общностей h_j:

метод наибольшей корреляции. На главной диагонали с положительным знаком записывается наибольший по величине коэффициент корреляции;

метод Барта. По каждому столбцу матрицы *R* вначале находят среднее значение коэффициентов корреляции $\overline{r_j}$, затем, если $\overline{r_j}$, сравнительно велико, за общность принимается значение, которое несколько выше наибольшего в столбце коэффициента корреляции и, если $\overline{r_j}$ – сравнительно малое значение, общность будет несколько меньше наибольшего в столбце коэффициента корреляции;

метод триад. Общности для каждого j -го столбца *R* вычисляют по формуле:

$$h_j^2 = \frac{r_{ik} r_{il}}{r_{kl}},$$

где r_{ik} r_{il} – коэффициенты корреляции, наибольшие в столбце;

метод малого центроида. Для каждой переменной *j* строится корреляционная матрица размерности 4x4. Включая саму переменную в эту матрицу, записывают оценки корреляции трех других переменных, особенно тесно связанных с первой. По данным малой матрицы корреляций и рассчитывают общности:

$$h_j^2 = \frac{\left(\sum r_{i1}\right)^2}{\sum r_{ij}},$$

где $\sum r_{i1}$ – сумма элементов первого столбца; $\sum r_{ij}$ – сумма всех элементов матрицы 4x4.

После определения редуцированной матрицы находят матрицу факторных нагрузок и по ее данным интерпретируют латентные факторы. Наилучшие решения находят при помощи современных методов факторного анализа: главных факторов, максимального правдоподобия и др. В общем случае выделенные факторы не обязательно ортогональны и тогда векторы (столбцы) матрицы факторных нагрузок будут линейно-зависимыми.

4.2 Моделирование значений наблюдаемых переменных на основе

выделенных латентных факторов

Методы факторного анализа при всем их многообразии имеют общий алгоритм решения. Начинаясь построением матрицы исходных данных *X,* этот алгоритм завершается получением матриц факторного отображения (факторных нагрузок) и значений факторов (факторных весов) – *A* и *F.* С учетом принятых обозначении где *n* – число наблюдений, *m* – число аналитических признаков *X,* *r* – число значимых обобщенных признаков (латентных факторов), на схеме (рис. 10) показана размерность матриц данных для каждого алгоритмического шага.

Первые шаги алгоритма 1–3 не вызывают каких-либо затруднений. Переход от матрицы исходных данных *X* к матрице стандартизованных данных *Z* осуществляется после пересчета всех элементов.

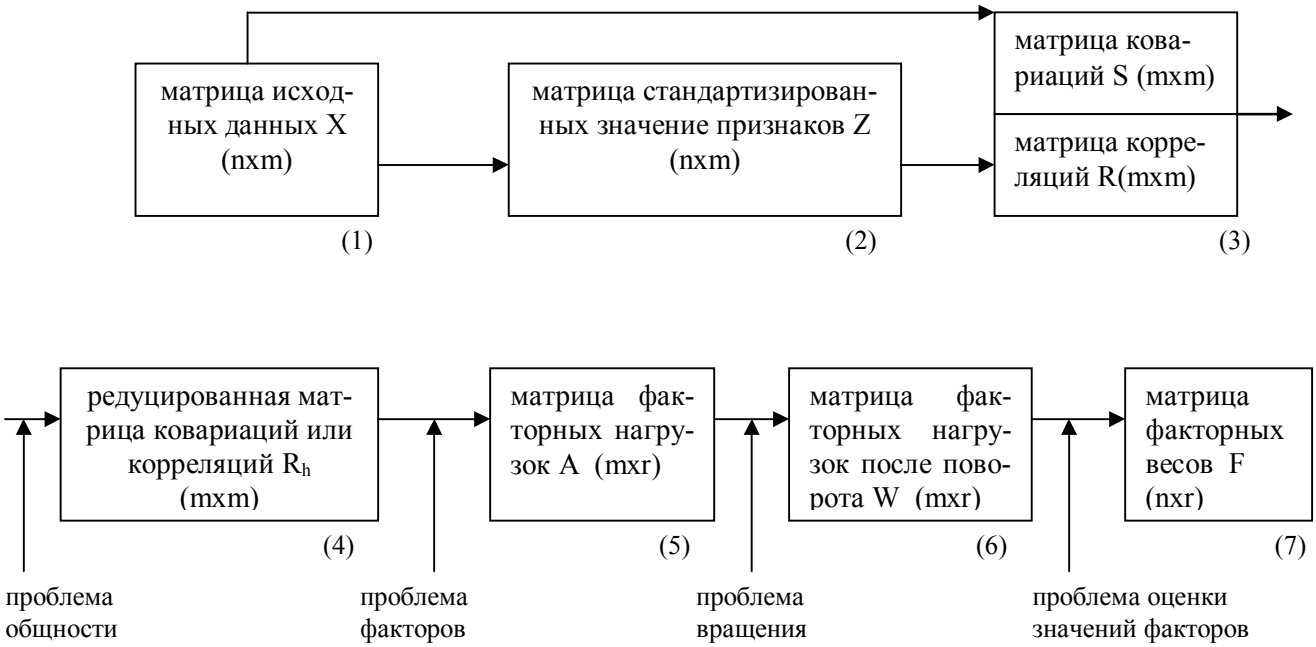

Рис. 10

На следующем шаге простым перемножением скаляра 1/n и матриц Z^T и Z получаем матрицу парных корреляций: $R = 1/nZ^T Z$.

Шаг 2 может быть опущен и тогда последующее факторное решение находят не по матрице корреляций, а по матрице ковариаций, но тогда анализируемые признаки должны иметь одни и те же единицы измерения.

Выполнение четвертого шага алгоритма обуславливается решением первой проблемы – построения редуцированной матрицы корреляций с общностями на главной диагонали.

Вторая проблема возникает на этапе построения матрицы факторных нагрузок *A* и заключается в выборе оптимального метода для поиска весовых коэффициентов *a* элементов матрицы *A*.

Выполнение шага 6 алгоритма и решение проблемы вращения пространства общих факторов не обязательно.

На последнем этапе алгоритма необходимо получить значения каждого из выделенных факторов для каждого индивидуального объекта исследования, т.е. матрицу факторных весов.

На основе исходных данных в матрице значений Y и матрицы *A* возможно получить оценки элементов матрицы *F*. В зависимости от решаемой задачи по этим оценкам можно судить о каждом объекте исследования по *m* общим факторам.

Для уяснения методики приступим к оценке *F* в методе главных компонент.

$Y=AF.$

Y имеет размерность (nxN); порядок A равен n, а F - (nxN). Поскольку при извлечении всех главных компонент матрица A квадратная, то задача получения матрицы F не вызывает затруднений, если матрица R имеет ранг, равный n. Умножим обе части равенства слева на A^{-1}, получим

$$F = A^{-1}Y$$

По этой формуле получаются точно и однозначно индивидуальные значения главных компонент для каждого объекта исследования.

Чаще всего извлекаются не все главные компоненты, а только (m<n), поэтому матрица A не квадратная, а значит не имеет обратной матрицы. В этом случае для нахождения матрицы факторных весов необходимо в первоначальном равенстве обе части слева умножить на $(A^TA)^{-1}A^T$, получим:

$\left(A^T A\right)^{-1} A^T Y = \left(A^T A\right)^{-1} A^T AF = A^{-1}\left(A^T\right)^{-1} A^T AF = A^{-1} EAF = F$, т.е.

$F = \left(A^T A\right)^{-1} A^T Y$

В этом выражении не надо обращать матрицу А. Если А не квадратная матрица, то $\left(A^T A\right)$ будет квадратной порядка m.

Дать точное определение индивидуальных значений факторов, как в случае выделения всех факторов не возможно. Поскольку задача не решается однозначно, то можно методом наименьших квадратов получить оценки индивидуальных значений общих факторов. Удобно обратиться к методу регрессионного анализа, когда имеется одна зависимая нормированная переменная и n независимых переменных, которые связаны между собой линейно.

$y_{0i} = \beta_1 y_{1i} + \beta_2 y_{2i} + \ldots + \beta_n y_{ni} + e_i$

Коэффициенты β_j выбираются таким образом, чтобы сумма квадратов ошибок оценок e^2 была минимальной.

Произведение корреляционной матрицы R на вектор-столбец коэффициентов регрессии β равен вектору-столбцу коэффициентов корреляции между оценками нормированных значений зависимой переменной и всеми исходными признаками V,т.е. $V = R\beta$, $\beta = R^{-1}V$, $\beta^T = (R^{-1}V)^T = V^T R^{-1}$

Матрица оценок индивидуальных значений факторов может быть определена по формуле: $F = \beta^T Y = V^T R^{-1} Y$,

где элементами матрицы V являются коэффициенты корреляции между переменными и факторами (матрица факторных нагрузок), R – матрица коэффициентов корреляции между переменными, Y – матрица исходных данных.

Как правило, проведение факторного анализа заканчивается оценкой индивидуальных значений факторов для каждого объекта исследования.

4.3 Методы главных компонент и главных факторов, вращение факторов

Потребность во вращении возникает, когда пространственное расположение общих факторов F_r нелогично или трудно поддается интерпретации. Воз-

можность появления нелогичных первых результатов анализа объясняется не определяемым четко и не задаваемым положением факторных осей в пространстве, или, отсутствием изначально какой-либо пространственной привязки для осей F_r.

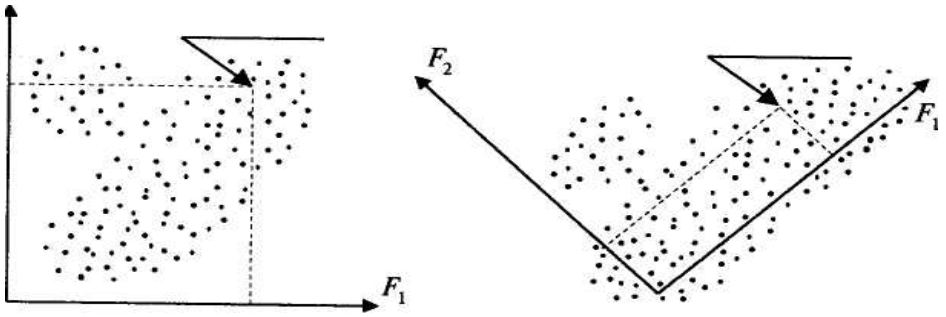

Рис. 11

На рисунке 11 показаны два различных положения в пространстве факторных осей *{F_1 и F_2).* Легко заметить, что изменение положения F_1 и F_2 одновременно приводит к изменению координат исходных признаков X_j. Цель поворота – преобразование координат (факторных нагрузок) таким образом, чтобы факторобразующие признаки имели наибольшие нагрузки, близкие к единице, а остальные признаки – минимальные значения, близкие к нулю, т.е. добиваются экономичного описания данных.

Повороты осей могут быть ортогональными и косоугольными. Предпочтительно, хотя и более трудно выполнимо и интерпретируемо, косоугольное вращение, при этом, значительно повышаются возможности оптимального отображения сгущений признаков в пространстве R^F.

На рисунке 12 ось F' после поворота F, очевидно, займет более рациональное положение, но из-за жесткости осевой конструкции положение F_2 удаляется от оптимального; на рис. б косоугольным вращением приходят к оптимизации положения сразу обеих осей F_1 и F_2

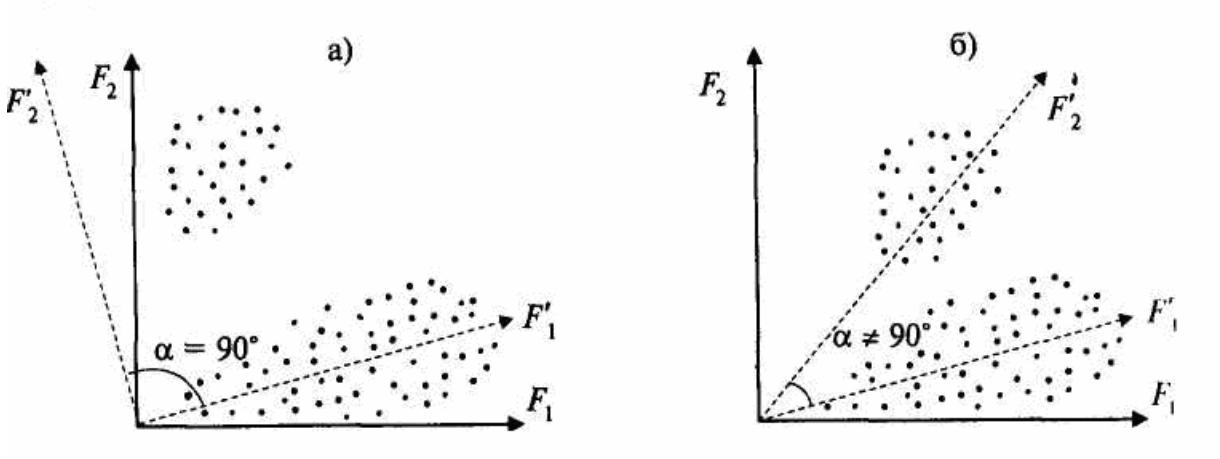

Рис. 12

Вращение пространства общих факторов F_r не изменяет величин общностей h и по-прежнему $AA^T = R^+$, или $ACA^T = R^+$ при $R^+ \rightarrow R$.

Рассмотрим особенности наиболее часто применяющихся методов главных компонент и главных факторов, которые имеют много общего.

Метод главных компонент (Г. Хотеллинг) строго говоря, не относится к факторному анализу, хотя он имеет с ним много общего. Специфическим является, во-первых, то, что в ходе вычислительных процедур одновременно получают все главные компоненты и их число первоначально равно числу элементарных признаков; во-вторых, постулируется возможность полного разложения дисперсии элементарных признаков, другими словами, ее полное объяснение через латентные факторы (обобщенные признаки).

Метод главных факторов заключается в том, что дисперсия элементарных признаков здесь объясняется не в полном объеме, признается, что часть дисперсии остается нераспознанной как характерность. Факторы выделяются последовательно: первый, объясняющий наибольшую долю вариации элементарных признаков, затем второй, объясняющий меньшую, вторую после первого латентного фактора часть дисперсии, третий и т.д. Процесс выделения факторов может быть прерван на любом шаге, если принято решение о достаточности доли объясненной дисперсии элементарных признаков или с учетом интерпретируемости латентных факторов.

Алгоритм обоих методов начинается с получения матрицы парных коэф-

фициентов корреляции с единицами на главной диагонали. После этого определяются общности и получают редуцированную матрицу.

Для определения латентных факторов находят собственные числа и собственные векторы редуцированной корреляционной матрицы.

Если использовать метод главных компонент, то необходимо определить собственные векторы для каждого собственного числа. В методе главных факторов первоначально отбирают значимые собственные числа, количество которых соответствует количеству главных факторов. Для этого используют критерий Кайзера или критерий факторной осыпи.

1. Критерий Кайзера предлагает отобрать только факторы, с собственными значениями, большими *1*.

2. Критерий факторной осыпи является графическим методом (рис. 13), впервые предложенным Кэттелем (Cattell). Необходимо изобразить собственные значения, расположенные в убывающем порядке в виде графика (по оси абсцисс порядковый номер числа, по оси ординат его значение).

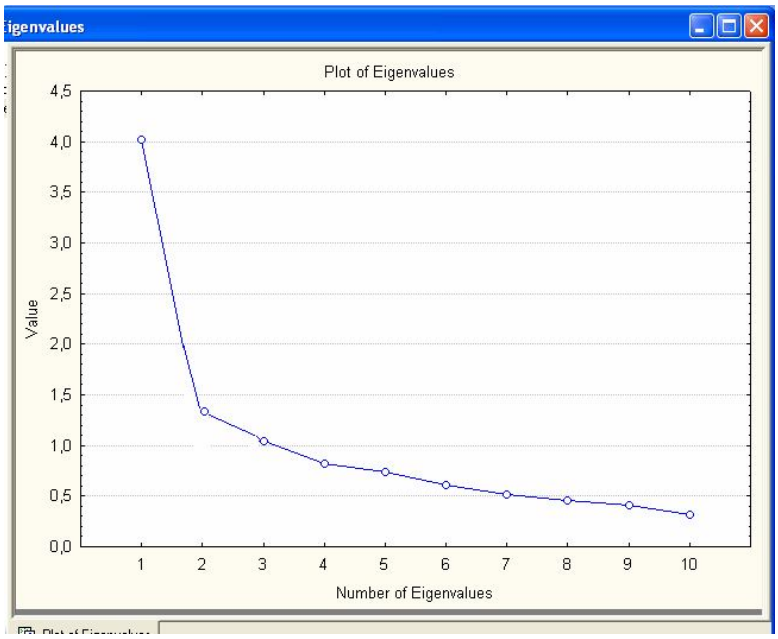

Рис. 13

Кэттель предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Значения выше этой точки определяют оптимальное количество факторов. Предполагается, что

справа от этой точки находится только «факториальная осыпь» – «осыпь» является геологическим термином, обозначающим обломки горных пород, скапливающиеся в нижней части скалистого склона. В соответствии с этим критерием можно оставить в приведенном примере 2 или 3 фактора.

Первый критерий (*критерий Кайзера*) иногда сохраняет слишком много факторов, в то время как второй критерий (*критерий каменистой осыпи*) иногда сохраняет слишком мало факторов; однако оба критерия вполне хороши при нормальных условиях, когда имеется относительно небольшое число факторов и много переменных. Обычно исследуется несколько решений с большим или меньшим числом факторов, и затем выбирается одно наиболее интерпретируемое.

Независимо от метода для выделенных собственных чисел определяются собственные векторы матрицы. Нормированные координаты собственных векторов, умноженные на весовой коэффициент, который равен корню из соответствующего собственного числа, являются столбцами матрицы факторных нагрузок.

Алгоритмы факторного анализа отличаются, трудоемкостью, их полное выполнение возможно при условии использования технических средств.

Пример 1

При переходе детей из начальной школы в классы среднего звена решено организовать несколько профильных классов (математический, гуманитарный, спортивный и др.). В связи с этим социологу поручено разработать систему тестов для младших школьников, с помощью которых можно определить способности ребенка. Фрагмент результатов тестов приведен в таблице 18.

Таблица 18

	Рост	Вес	Память на числа	Логическое мышление	Техника чтения	Фантазия	Прыжки в длину
Ира	150	35	7	1	180	0	70
Катя	156	34	6	0	156	1	120
Маша	168	32	5	1	189	0	140

Даша	160	39	2	0	190	1	80
Юля	162	45	4	0	200	1	90
Таня	159	30	5	0	230	0	100
Кира	145	28	9	1	250	1	120
Оля	150	32	7	1	200	0	110
Поля	160	25	7	1	180	0	100

Сжать матрицу исходных данных путем выделения оптимального количества факторов.

Решение На первом этапе необходимо стандартизировать исходные данные. Матрица стандартизированных данных приведена в таблице 19.

Таблица 19.

	рост	вес	Память на числа	Логическое мышление	Техника чтения	фантазия	Прыжки в длину
Ира	-0,93	0,28	0,60	0,84	-0,61	-0,84	-1,53
Катя	-0,09	0,11	0,11	-1,05	-1,47	1,05	0,76
Маша	1,58	-0,22	-0,38	0,84	-0,29	-0,84	1,68
Даша	0,47	0,95	-1,84	-1,05	-0,26	1,05	-1,07
Юля	0,74	1,96	-0,87	-1,05	0,10	1,05	-0,61
Таня	0,33	-0,56	-0,38	-1,05	1,17	-0,84	-0,15
Кира	-1,63	-0,90	1,57	0,84	1,88	1,05	0,76
Оля	-0,93	-0,22	0,60	0,84	0,10	-0,84	0,31
Поля	0,47	-1,40	0,60	0,84	-0,61	-0,84	-0,15

Далее по матрице стандартизированных данных надо определить коэффициенты корреляции между всеми переменными (табл. 20).

Таблица 20

	Рост	Вес	Память на числа	Логическое мышление	Техника чтения	Фантазия	Прыжки в длину
Рост	1,00	0,26	-0,70	-0,34	-0,36	-0,12	0,18
Вес	0,26	1,00	-0,67	-0,58	-0,23	0,50	-0,42
Память	-0,70	-0,67	1,00	0,71	0,27	-0,24	0,30
Логическое мышление	-0,34	-0,58	0,71	1,00	0,11	-0,55	0,25
Техника чтения	-0,36	-0,23	0,27	0,11	1,00	0,06	0,12
Фантазия	-0,12	0,50	-0,24	-0,55	0,06	1,00	-0,04
Прыжки в	0,18	-0,42	0,30	0,25	0,12	-0,04	1,00

длину							

Для определения оптимального количества латентных факторов необходимо найти собственные числа корреляционной матрицы путем решения уравнения: $|R - \lambda E| = 0$, определить сколько % дисперсии объясняет каждое собственное число: $\frac{\lambda_i}{\sum \lambda} 100\%$ и вычислить накопленный процент (табл. 21).

Таблица 21

Собственные числа	% объясненной дисперсии	Накопленный %
3,071116	43,87308	43,8731
1,484121	21,20173	65,0748
1,072478	15,32112	80,3959
0,738453	10,54932	90,9453
0,373828	5,34040	96,2857
0,172748	2,46783	98,7535
0,087256	1,24652	100,0000

Согласно критерию Кайзера значимыми являются три фактора (с собственными числами больше 1), они суммарно объясняют чуть более 80% всей дисперсии.

Для того чтобы воспользоваться критерием Кеттеля, необходимо изобразить собственные числа в системе координат. По оси OX располагаем порядковый номер собственного числа, а по оси OY – значения собственных чисел (рис. 14).

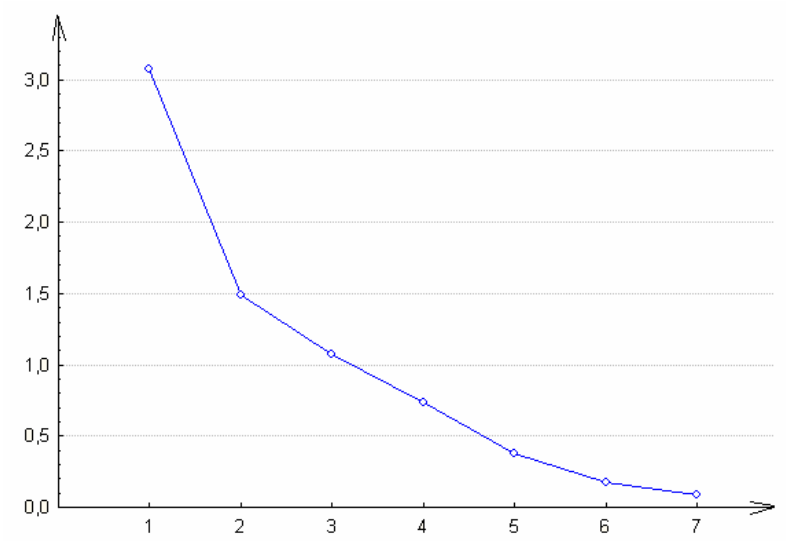

Рис. 14

Согласно графику факторной осыпью можно считать собственные числа, начиная с пятого, так как именно с этой точки график начинает плавно приближаться к оси ОХ. Причем четыре оставшихся собственных числа объясняют всю дисперсию на 90,9 %.

Итак, оптимальное количество факторов данной задачи 3 или 4, более точный результат можно получить путем возможности интерпретации факторов.

Коэффициентами матрицы факторных нагрузок являются нормированные собственные векторы корреляционной матрицы, соответствующие выделенным раннее собственным числам.

Собственные векторы определяются путем решения систем линейных однородных уравнений вида $(R-\lambda E)X=0$, а их нормирование производится по формуле $x_i^н=\frac{x_i}{\sqrt{x_1^2+x_2^2+...+x_n^2}}$. Кроме того, после нормирования каждую координату вектора необходимо умножить на весовой коэффициент, в качестве которого выступает $\sqrt{\lambda}$.

Итак, для найденных на предыдущем занятии собственных чисел λ_1=3,071116; λ_2=1,484121; λ_3=1,072478; λ_4=0,738453 нормированные взвешенные собственные векторы равны:

x_1=(-0,55; -0,44; 0,91; 0,84; 0,36; -0,5; 0,39);

x_2=(-0,24; 0,23; 0,23; -0,20; 0,52; 0,63; -0,36);

x_3=(0,74; 0,12; 0,03; 0,22; -0,38; -0,41; -0,8);

x_4=(-0,19; 0,07; 0,25; 0,08; -0,67; 0,37; 0,22).

Располагая координаты полученных собственных векторов по столбцам (табл. 22), а данные переменные по строке получим матрицу факторных нагрузок.

Таблица 22

	Фактор 1	Фактор 2	Фактор 3	Фактор 4
Рост	-0,55	-0,24	**0,74**	-0,19

Вес	-0,44	0,23	0,12	0,07
Память	**0,91**	0,23	0,03	0,25
Логическое мышление	**0,84**	-0,20	0,22	0,08
Техника чтения	0,36	**0,52**	-0,38	-0,67
Фантазия	-0,50	**0,63**	-0,41	0,37
Прыжки в длину	0,39	-0,36	**-0,80**	0,22

В матрице факторных нагрузок содержатся коэффициенты корреляции между выделенными факторами и исходными переменными, анализ которых показывает, что первый фактор сильнее всего зависит от тестов на память и логическое мышление, т.е. может отвечать за математические способности школьников; второй фактор – гуманитарной направленности; третий – спортивной, а четвертый фактор не поддается интерпретации. В связи с этим в матрице факторных нагрузок оставляем три фактора, а четвертый отбрасываем.

Элементы матрицы факторных весов являются значениями выделенных факторов для каждого испытуемого. Матрица факторных весов для данной задачи приведена в таблице 23.

Таблица 23

факторы	математический	гуманитарный	спортивный
Марина	-0,74	-0,94	-1,45
Катя	1,31	-0,70	1,02
Маша	-1,03	1,44	1,60
Даша	0,26	0,90	-1,09
Юля	0,27	0,70	-0,38
Таня	1,79	0,28	-0,15
Кира	-0,80	-1,19	0,55
Оля	-0,29	-1,15	0,52
Поля	-0,78	0,66	-0,61

Анализируя матрицу можно сказать, что все выделенные способности у Марины ниже среднего (факторные весы <0), но преобладают математические.

В математический класс целесообразно направить Катю и Таню; в гуманитарный – Дашу, Юлю, Полю; а в спортивный – Машу, Киру, Олю.

<u>Пример 2</u> По данным о численности (x_1) и фонде зарплаты (x_2) пяти(n=5) строительных организаций провести компонентный анализ.

$$X=\begin{pmatrix}3 & 4\\ 6 & 5\\ 8 & 9\\ 2 & 3\\ 7 & 6\end{pmatrix}.$$

<u>Решение.</u> Рассчитаем выборочные характеристики переменных x_1 и x_2:

$\overline{x_1}$=5,2; S_1=2,315; $\overline{x_2}$=5,4; S_2=2,059, $\overline{x_1x_2}$=32,4.

Выборочный коэффициент корреляции равен:

$$r=\frac{\overline{x_1x_2}-\overline{x_1}\cdot\overline{x_2}}{S_1S_2}=\frac{32{,}4-5{,}2\cdot 5{,}4}{2{,}315\cdot 2{,}059}=0{,}906.$$

Преобразуем матрицу X в матрицу нормированных значений Z, с элементами: $z_{ij}=\dfrac{x_{ij}-\overline{x_j}}{S_j}$, где i=1,2,3,4,5; j=1,2.

$$Z=\begin{pmatrix}-0{,}950 & -0{,}680\\ 0{,}346 & -0{,}194\\ 1{,}210 & 1{,}748\\ -1{,}382 & -1{,}166\\ 0{,}778 & 0{,}291\end{pmatrix}.$$

Матрица парных коэффициентов корреляции имеет вид: $R=\begin{pmatrix}1 & 0{,}906\\ 0{,}906 & 1\end{pmatrix}$.

Для определения собственных значений матрицы R, рассмотрим характеристическое уравнение:

$$|\lambda E-R|=\left|\begin{pmatrix}\lambda & 0\\ 0 & \lambda\end{pmatrix}-\begin{pmatrix}1 & 0{,}906\\ 0{,}906 & 1\end{pmatrix}\right|=\begin{vmatrix}\lambda-1 & 0{,}906\\ 0{,}906 & \lambda-1\end{vmatrix}=0.$$

$(\lambda - 1)^2 - (0{,}906)^2=0$, $\lambda - 1=\pm 0{,}906$.

Т.к. по условию компонентного анализа $\lambda_1>\lambda_2$, то λ_1=1,906, λ_2=0,094, где λ_1, λ_2 соответственно дисперсии и вклад 1-й и 2-й главных компонент в суммарную дисперсию, равную $\lambda_1+\lambda_2$=k=2.

Относительный вклад компонент в суммарную дисперсию равен:

$$\frac{\lambda_1}{k}\cdot 100\% = \frac{1{,}906}{2}\cdot 100\% = 95{,}3\%;\ \frac{\lambda_2}{k}\cdot 100\% = \frac{0{,}094}{2}\cdot 100\% = 4{,}7\%.$$

Таким образом, $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = \begin{pmatrix} 1{,}906 & 0 \\ 0 & 0{,}094 \end{pmatrix}$.

Определим матрицу собственных векторов из уравнения (R-λE)V=0.

Откуда собственный вектор V_1 находим из условия: $\begin{pmatrix} 1-\lambda_1 & r \\ r & 1-\lambda_1 \end{pmatrix}\begin{pmatrix} V_{11} \\ V_{21} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, где $V_1 = \begin{pmatrix} V_{11} \\ V_{21} \end{pmatrix}$.

Подставляя полученные значения имеем: $\begin{pmatrix} -0{,}906 & 0{,}906 \\ 0{,}906 & 0{,}906 \end{pmatrix}\begin{pmatrix} V_{11} \\ V_{21} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

Откуда, $-0{,}906v_{11}+0{,}906v_{21}=0$ или $v_{11}=v_{21}=1$, т.е. $V_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Нормированный собственный вектор, соответствующий λ_1, равен: $U_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$.

Собственный вектор V_2 находим из условия:

$$\begin{pmatrix} 0{,}906 & 0{,}906 \\ 0{,}906 & 0{,}906 \end{pmatrix}\begin{pmatrix} V_{12} \\ V_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Откуда, $0{,}906v_{12}+0{,}906v_{22}=0$ или $-v_{12}=v_{22}=1$, т.е. $V_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$.

Нормированный собственный вектор, соответствующий λ_2, равен: $U_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$.

Тогда нормированная матрица собственных векторов имеет вид:

$$U = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 0{,}707 & -0{,}707 \\ 0{,}707 & 0{,}707 \end{pmatrix}.$$

Матрицу факторных нагрузок найдем по формуле:

$A = U\Lambda^{1/2}$, где $\Lambda^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{pmatrix}$.

Подставив полученные значения, получим:

$$A = \begin{pmatrix} 0{,}707 & -0{,}707 \\ 0{,}707 & 0{,}707 \end{pmatrix} \times \begin{pmatrix} 1{,}3807 & 0 \\ 0 & 0{,}3063 \end{pmatrix} = \begin{pmatrix} 0{,}9763 & -0{,}2166 \\ 0{,}9763 & 0{,}2166 \end{pmatrix}.$$

Матрицу факторных нагрузок используют для интерпретации главных компонент, т.к. элементы матрицы $a_{jv}=r_{jv}$ характеризуют тесноту связи между x_j-м признаком и f_v главной компонентой. В нашем примере первая главная компонента тесно связана с показателями X_1 и X_2, f_1 – характеризует размер предприятия.

Матрицу значений главных компонент F можно получить по формуле: $F = Z(A^T)^{-1}$.

Так как $A^T = \begin{pmatrix} 0{,}9763 & 0{,}9763 \\ -0{,}2166 & 0{,}2166 \end{pmatrix}$,

то $(A^T)^{-1} = \frac{1}{0{,}4229}\begin{pmatrix} 0{,}2166 & 0{,}2166 \\ -0{,}9763 & 0{,}9763 \end{pmatrix} = \begin{pmatrix} 0{,}5121 & 0{,}5121 \\ -2{,}3084 & 2{,}3084 \end{pmatrix}$.

Тогда $F = \begin{pmatrix} -0{,}950 & -0{,}680 \\ 0{,}346 & -0{,}194 \\ 1{,}210 & 1{,}748 \\ -1{,}382 & -1{,}166 \\ 0{,}778 & 0{,}291 \end{pmatrix} \times \begin{pmatrix} 0{,}5121 & 0{,}5121 \\ -2{,}3084 & 2{,}3084 \end{pmatrix} = \begin{pmatrix} 1{,}0832 & -2{,}0562 \\ 0{,}6250 & -0{,}2706 \\ -3{,}4154 & 4{,}6547 \\ 1{,}9839 & -3{,}3993 \\ -0{,}2733 & 1{,}0702 \end{pmatrix}$

Матрица F, которую мы получили, характеризует пять строительных организаций в пространстве главных компонент. Ее можно использовать в задачах классификации и регрессионного анализа. Например, классификация организаций по первой главной компоненте позволяет их ранжировать следующим образом: 3, 5, 2, 1, 4, что согласуется с матрицей X.

Задание для самостоятельной работы

Определите удовлетворенность, какой стороной жизни преобладает у разных респондентов. Пусть в эксперименте участвовало 10 человек, которые отвечали на вопросы, представленные в таблице 24.

Таблица 24.

Возраст	Образование	Стаж непрерывной работы	Работа по специальности	Зарплата, тыс.руб	Кол-во человек в семье	Жилплощадь	Наличие хобби	Посещение учреждение вне работы

Образование: 2 – высшее, 1 – специальное, 0 – среднее.

Работа по специальности: 1 – да, 0 – нет.

Хобби: 1 – да, 0 – нет.

Посещение учреждение вне работы: 3 – часто, 2 – иногда, 1 – нет.

Проведите стандартизацию данных и определите оптимальное количество факторов, рассчитайте матрицы факторных нагрузок и факторных весов, сделайте вывод.

5 ДИСКРИМИНАНТНЫЙ АНАЛИЗ

5.1 Постановка задачи

Дискриминантный анализ является разделом многомерного статистического анализа, который позволяет изучать различия между двумя и более группами объектов по нескольким переменным одновременно.

При интерпретации нужно ответить на вопрос: возможно ли, используя данный набор переменных, отличить одну группу от другой, насколько хорошо эти переменные помогают провести дискриминацию и какие из них наиболее информативны.

Методы классификации связаны с получением одной или нескольких функций, обеспечивающих возможность отнесения данного объекта к одной из групп. Эти функции называются классифицирующими и зависят от значений

переменных таким образом, что появляется возможность отнести каждый объект к одной из групп.

Задачи дискриминантного анализа можно разделить на три типа. Задачи первого типа позволяют на основе некоторой информации найти функцию, позволяющую поставить в соответствие новым объектам характерные для нихвыводы. Построение такой функции и составляет задачу дискриминации. Второй тип задачи относится к ситуации, когда признаки принадлежности объекта к той или иной группе потеряны, и их нужно восстановить. Задачи третьего типа связаны с предсказанием будущих событий на основании имеющихся данных.

Дискриминантный анализ общий термин, относящийся к нескольким тесно связанным статистическим процедурам. Эти процедуры можно разделить на методы интерпретации межгрупповых различий – *дискриминации* и методы *классификации* наблюдений по группам.

5.2 Построение дискриминантных функций

Основной целью дискриминации является нахождение такой линейной комбинации переменных (в дальнейшем эти переменные будем называть дискриминантными переменными), которая бы оптимально разделила рассматриваемые группы.

Введем следующие обозначения:

g – число классов;

p – число дискриминантных переменных;

n_k – число наблюдений в *k*–й группе;

n – общее число наблюдений по всем группам;

x_{ikm} – величина дискриминантной переменной *i* для *m*–го наблюдения в *k*–й группе;

$\bar{x}_{ik}$ – средняя величина переменной *i* в *k*–й группе;

$\bar{x}_i$ – среднее значение переменной i по всем группам;

Линейная функция $d_{km} = \beta_0 + \beta_1 x_{1km} + \ldots + \beta_p x_{pkm}$, $m = 1,\ldots n$; $k = 1,\ldots g$ называется *канонической дискриминантной функцией* с неизвестными коэффициен-

тами β_i, d_{km} –значение дискриминантной функции для *m*–го объекта в группе *k*

Коэффициенты β_i первой канонической дискриминантной функции выбираются таким образом, чтобы центроиды различных групп как можно больше отличались друг от друга. Коэффициенты второй группы выбираются также, но при этом налагается дополнительное условие, чтобы значения второй функции были некоррелированы со значениями первой. Аналогично определяются и другие функции. Отсюда следует, что любая каноническая дискриминантная функция имеет нулевую внутригрупповую корреляцию с остальными. Если число групп равно *g*, то число канонических дискриминантных функций будет на единицу меньше числа групп. С геометрической точки зрения дискриминантные функции определяют гиперповерхности в *p*-мерном пространстве. В частном случае при *p*=2 она является прямой, а при *p*=3 плоскостью

На практике полезно иметь одну, две или же три дискриминантных функций. Тогда графическое изображениее объектов будет представлено в одно–, двух– и трехмерных пространствах. Такое представление особенно полезно в случае, когда число дискриминантных переменных p велико по сравнению с числом групп g.

Для получения коэффициентов β_i канонической дискриминантной функции нужен статистический критерий различения групп. Классификация переменных будет осуществляться тем лучше, чем меньше рассеяние точек относительно центроида внутри группы и чем больше расстояние между центроидами групп. Один из методов поиска наилучшей дискриминации данных заключается в нахождении такой канонической дискриминантной функции d, которая бы максимизировала отношение межгрупповой вариации к внутригрупповой .

$$\lambda = \frac{B}{W}$$

где B – межгрупповая, а W – внутригрупповая матрицы рассеяния наблюдаемых переменных от средних. Иногда вместо W используют матрицу рассеяния T объединенных данных.

Рассмотрим задачу максимизации отношения для λ когда имеются g групп. Оценим информацию, характеризующую степень различия между объектами по всему пространству точек, определяемому переменными групп. Для этого вычислим матрицу рассеяния T, которая равна сумме квадратов отклонений и попарных произведений наблюдений от общих средних по каждой переменной. Элементы матрицы T определяются выражением

$$t_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n} (x_{ikm} - \bar{x}_i)(x_{jkm} - \bar{x}_j) \qquad \bar{x}_i = (1/n)\sum_{k=1}^{g} n_i \bar{x}_{ik}, i = 1, \ldots, p$$

$$\bar{x}_{ik} = (1/n_i)\sum_{m=1}^{n_k} x_{ikm}, i = 1, \ldots, p; k = 1, \ldots, g$$

Матрица T содержит полную информацию о распределении точек по пространству переменных. Диагональные элементы представляют собой сумму квадратов отклонений от общего среднего и показывают, как ведут себя наблюдения по отдельно взятой переменной. Внедиагональные элементы равны сумме произведений отклонений по одной переменной на отклонения по другой.

Если разделить матрицу T на (n-1), то получим ковариационную матрицу. Для проверки условия линейной независимости переменных полезно рассмотреть вместо T нормированную корреляционную матрицу.

Для измерения степени разброса объектов внутри групп рассмотрим матрицу W, которая отличается от T только тем, что ее элементы определяются векторами средних для отдельных групп, а не вектором средних для общих данных. Элементы внутригруппового рассеяния определятся выражением

$$w_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (x_{ikm} - \bar{x}_{ik})(x_{jkm} - \bar{x}_{jk})$$

Если разделить каждый элемент матрицы W на (n–g), то получим оценку ковариационной матрицы внутригрупповых данных.

Когда центроиды различных групп совпадают, то элементы матриц T и W будут равны. Если же центроиды групп различные, то разница B=T-W будет определять межгрупповую сумму квадратов отклонений и попарных произве-

дений. Если расположение групп в пространстве различается (т.е. их центроиды не совпадают), то степень разброса наблюдений внутри групп будет меньше межгруппового разброса.

Матрицы W и B содержат всю основную информацию о зависимости внутри групп и между группами. Для лучшего разделения наблюдений на группы нужно подобрать коэффициенты дискриминантной функции из условия максимизации отношения межгрупповой матрицы рассеяния к внутригрупповой матрице рассеяния при условии ортогональности дискриминантных плоскостей. Тогда нахождение коэффициентов дискриминантных функций сводится к решению задачи о собственных значениях и векторах

Пусть $\lambda_1 \geq \ldots \lambda_p$ и $v_1 \ldots v_p$ соответственно собственные значения и векторы. Тогда $\lambda = \frac{\sum_k b_{jk} v_j v_k}{\sum_k w_{jk} v_j v_k}$, что влечет равенство $\sum_k (b_{jk} - \lambda w_{jk}) v_k = 0$, или в матричной записи $(B - \lambda W)v = 0$.

Решение уравнения $|B - \lambda W| = 0$ позволяет нам определить компоненты собственных векторов, соответствующих дискриминантным функциям.

Каждое решение, которое имеет свое собственное значение и собственный вектор, соответствует одной дискриминантной функции. Координаты собственного вектора можно использовать в качестве коэффициентов дискриминантной функции. Однако при таком подходе начало координат не будет совпадать с главным центроидом. Для того, чтобы начало координат совпало с главным центроидом нужно нормировать координаты собственного вектора.

$$\beta_j = v_i \sqrt{n - g}\,,\ \beta_0 = -\sum_{j=1}^{p} \beta_i \overline{x_i}\,.$$

Нормированные коэффициенты полученые по нестандартизованным исходным данным, называются *нестандартизованными.* Нормированные коэффициенты приводят к таким дискриминантным значениям, единицей измерения которых является стандартное квадратичное отклонение. При таком подходе каждая ось в преобразованном пространстве сжимается или растягивается та-

ким образом, что соответствующее дискриминантное значение для данного объекта представляет собой число стандартных отклонений точки от главного центроида.

Стандартизованные коэффициенты можно получить двумя способами:

1. стандартизировать исходные данные;

2. преобразовать нестандартизованные коэффициенты к стандартизованной форме по формуле:

$$c_j = \beta_j \sqrt{\frac{w_{ji}}{n-g}},$$

Стандартизованные коэффициенты полезно применять для уменьшения размерности исходного признакового пространства переменных. Если абсолютная величина коэффициента для данной переменной для всех дискриминантных функций мала, то эту переменную можно исключить, тем самым сократив число переменных.

Стандартизованные коэффициенты показывают вклад переменных в значение дискриминантной функции. Если две переменные сильно коррелированы, то их стандартизованные коэффициенты могут быть меньше по сравнению с теми случаями, когда используется только одна из этих переменных. Такое распределение величины стандартизованного коэффициента объясняется тем, что при их вычислении учитывается влияние всех переменных.

Общее число дискриминантных функций не превышает числа дискриминантных переменных и, по крайней мере, на 1 меньше числа групп. Степень разделения выборочных групп зависит от величины собственных чисел: чем больше собственное число, тем сильнее разделение. Наибольшей разделительной способностью обладает первая дискриминантная функция, соответствующая наибольшему собственному числу вторая обеспечивает максимальное различение после первой и т.д. Различительную способность *i*–ой функции оценивают по относительной величине в процентах собственного числа от суммы всех собственных чисел.

Коэффициент канонической корреляции. Другой характеристикой, позво-

ляющей оценить полезность дискриминантной функции является коэффициент канонической корреляции r_i. Каноническая корреляция является мерой связи между двумя множествами переменных. Максимальная величина этого коэффициента равна 1. Будем считать, что группы составляют одно множество, а другое множество образуют дискриминантные переменные. Коэффициент канонической корреляции для *i*–ой дискриминантной функции определяется формулой:

$$r_j = \sqrt{\frac{\lambda_i}{1+\lambda_j}}.$$

Чем больше величина r_i, тем лучше разделительная способность дискриминантной функции.

После получения канонических дискриминантных функций при известной принадлежности объектов к тому или иному классу. решается задача предсказания класса, которому принадлежит некоторый случайно выбранный объект, т.е. задача классификации.

5.3 Построение классификации объектов

Классификация может проводиться с помощью апостериорной вероятности по Бейесовской схеме вычисления; с применением элементарных классифицирующих функций; с помощью расстояния Махалонобиса.

Классифицирующие функции имеют вид:

$$d_{ik} = b_{k0} + b_{k1}x_{i1} + \ldots + b_{kp}x_{ip} + \ln q_k, \;\; k = 1,\ldots,g$$

Объект относится к классу, у которого значение функции оказывается наибольшим. Коэффициенты классифицирующих функций удобнее вычислять по скалярным выражениям

$$b_{ki} = (n-g)\sum_{j=1}^{p}(w^{-1})_{ij}\bar{x}_{jk}, \;\; k = 1,\ldots,g$$

где b_{ki} – коэффициент для переменной *i* в выражении, соответствующему классу *k*,

$\left(w^{-1}\right)_{ij}$ – обратный элемент внутригрупповой матрицы сумм попарных

произведений W. Постоянный член находится по формуле

$$b_{k0} = -0{,}5 \cdot \sum_{j=1}^{p} b_{kj} \bar{x}_{jk},\ k = 1, \ldots, g$$

Выбор функций расстояния между объектами для классификации является наиболее очевидным способом введения меры сходства для векторов объектов, которые интерпретируются как точки в евклидовом пространстве. В качестве меры сходства можно использовать евклидово расстояние между объектами. Чем меньше расстояние между объектами, тем больше сходство. Однако в тех случаях, когда переменные коррелированы, измерены в разных единицах и имеют различные стандартные отклонения, трудно четко определить понятие "расстояния". В этом случае полезнее применить не евклидовое расстояние, а *выборочное расстояние Махаланобиса*

$$D^2(\mathbf{x} / G_k) = (n - g) \cdot \sum_{v=1}^{p} \sum_{j=1}^{p} (w^{-1})_{vj} (x_{iv} - \bar{x}_{vk})(x_{ij} - \bar{x}_{jk}),\ k = 1, \ldots, g$$

При использовании функции расстояния, объект относят к той группе, для которой расстояние наименьшее.

Пример

Методами кластерного анализа следующая совокупность разбита на три группы. Проверить достоверность полученного разбиения методами дискриминантного анализа, и классифицировать последние наблюдения.

В таблице представлены значения уровней сформированности компонент математической подготовки для 10 наблюдений и уровень качества математической подготовки.

№	исходные данные			Уровень качества
	Мотивация (МВ)	Знания, умения, навыки (КП)	Способность к исследовательской деятельности (ИР)	
1	1	1	2	1
2	2	2	1	1
3	1	2	1	1
4	2	3	2	3
5	1	3	3	3
6	1	2	2	2
7	3	3	1	2

8	2	3	1	2
9	2	1	1	1
10	2	1	2	1
11	3	3	2	?
12	1	2	3	?
13	1	3	2	?
14	3	2	1	?

Разбиения некоторой совокупности анализируемых объектов на классы осуществляется путем построения классифицирующей дискриминантной функции.

$$d_i = \beta_0 + \beta_1 \cdot МВ_i + \beta_2 \cdot КП_i + \beta_3 \cdot ИР_i, \qquad i = \overline{1,324}$$

β_j - неизвестные коэффициенты дискриминантной функции;

d_i - значение дискриминантной функции для *i*–го объекта в группе;

$МВ_i$, $КП_i$, $ИР_i$ – значения уровня сформированности соответствующих компонент математической подготовки для *i*-го наблюдения.

В качестве коэффициентов дискриминантной функции можно использовать нормированные координаты собственного вектора v_j отвечающего собственному числу λ_j, полученного из решения уравнения $|B - \lambda W| = 0$, где *B*- межгрупповая, а *W*- внутригрупповая матрицы рассеяния наблюдаемых переменных от средних.

B=T-W

где *T* матрица, характеризующая степень различия между объектами по всему пространству точек, определяемому переменными групп. Элементы матрицы *T* определяются выражением :

$$t_{ij} = \sum_{k=1}^{3} \sum_{m=1}^{324} (x_{ikm} - \overline{x_i})(x_{jkm} - \overline{x_j}),$$

где $\overline{x_i} = \frac{1}{n} \sum_{k=1}^{3} n_i \overline{x}_{ik}$ $\quad i = \overline{1,3}$ - среднее значение переменных (МВ, КП, ИР) по всем группам;

$\bar{x}_{ik} = \frac{1}{n_k}\sum_{m=1}^{n_k} x_{ikm}$, $i = \overline{1,3}$, $k = \overline{1,3}$ - среднее значение переменных (МВ, КП, ИР) в k-ой группе.

Элементы внутригруппового рассеяния определятся выражением:

$$w_{ij} = \sum_{k=1}^{3}\sum_{m=1}^{n_k}(x_{ikm} - \bar{x}_{ik})(x_{jkm} - \bar{x}_{jk})$$

Элементы матриц *T, B* и *W* приведены в таблице.

Матрица *T*			Матрица *W*			Матрица *B*		
164,14	9,07	21,59	147,58	-27,84	-8,73	16,63	36,91	30,32
9,07	147,22	40,09	-27,84	63,21	-30,19	36,91	84,02	70,28
21,59	40,09	119,44	-8,73	-30,19	59,89	30,32	70,28	59,55

Решая уравнение $|B - \lambda W| = 0$ относительно λ получаем, что собственные числа равны $\lambda_1 = 6{,}08$; $\lambda_2 = 0{,}01$; $\lambda_3 = 0{,}00003866$.

Степень разделения выборочных групп зависит от величины собственных чисел: чем больше собственное число, тем сильнее разделение. Наибольшей разделительной способностью обладает дискриминантная функция, соответствующая наибольшему собственному числу $\lambda_1 = 6{,}08$. Кроме того, только та функция, которая соответствует данному собственному числу является значимой, о чем свидетельствует анализ на основе Λ-статистики Wilks.

Λ-статистика вычисляется по формуле:

$$\Lambda = \prod_{i=k+1}^{g}(1/(1+\lambda_i))$$

где *k* число вычисленных функций. Чем меньше эта статистика, тем значимее соответствующая дискриминантная функция.

Величина

$$\chi^2 = -[n - ((p+g)/2) - 1]\ln\Lambda_k, \quad k = 0, 1, \ldots, g-1$$

имеет хи–квадрат распределение с $(p-k)(g-k-1)$ степенями свободы.

Значимость критерия χ^2 при *k*=0 подтверждает существование различий между группами. Кроме того, доказывает, что первая дискриминантная функция значима и имеет смысл ее вычислять.

Определяем первую дискриминантную функцию и проверяем значимость критерия при k=1. Если критерий значим, то вычисляем вторую дискриминантную функцию и продолжаем процесс до тех пор, пока не будет исчерпана вся значимая информация.

При κ=0 $\Lambda = \prod_{i=1}^{3}\left(\frac{1}{1+\lambda_i}\right) = \frac{1}{1+6{,}08}\cdot\frac{1}{1+0.01}\cdot\frac{1}{1+0.00003866} = 0.139$, тогда

$$\chi^2 = -(324-((3+3)/2)-1)\cdot\ln 0.139 = 629{,}48$$

$$\chi^2_{крит} = \begin{cases} 12{,}59 & при\ \alpha = 0{,}05 \\ 16{,}81 & при\ \alpha = 0{,}01 \end{cases}$$

Значимость критерия χ^2 очевидна.

При κ=1 $\Lambda = \prod_{i=2}^{3}\left(\frac{1}{1+\lambda_i}\right) = \frac{1}{1+0.01}\cdot\frac{1}{1+0.00003866} = 0{,}989$, тогда

$$\chi^2 = -(324-((3+3)/2)-1)\cdot\ln 0.989 = 3{,}22$$

$$\chi^2_{крит} = \begin{cases} 5{,}99 & при\ \alpha = 0{,}05 \\ 9{,}2 & при\ \alpha = 0{,}01 \end{cases}$$

В данном случае критерий не является значимым, что подтверждает необходимость лишь одной дискриминантной функции.

Определим координаты собственного вектора из решения уравнения $(B-\lambda W)\cdot V = O$.

Для $\lambda_1 = 6{,}08$ v_1=-0,317; v_2=-0,726; v_3=-0,61;

для $\lambda_2 = 0{,}01$ v_1=-0,75; v_2=-0,301; v_3=0,589.

Полученные координаты вектора можно использовать в качестве коэффициентов дискриминантной функции. Однако при таком подходе начало координат не будет совпадать с главным центроидом. Для того чтобы начало координат совпало с главным центроидом нужно нормировать координаты собственного вектора согласно формулам:

$$\beta_i = v_i\sqrt{324-3}, \quad \beta_0 = -\sum_{i=1}^{3}\beta_i\overline{x_i}$$

Для $\lambda_1 = 6{,}08$ $\beta_1 = -5{,}68$, $\beta_2 = -13{,}01$, $\beta_3 = -10{,}93$, $\beta_0 = 51{,}94$;

для $\lambda_2 = 0{,}01$ $\beta_1 = -13{,}44$, $\beta_2 = -5{,}39$, $\beta_3 = 10{,}55$, $\beta_0 = -15{,}65$

Дискриминантные функции имеют вид:

$$d_1 = 51{,}94 - 5{,}68 \cdot МВ_i - 13{,}01 \cdot КП_i - 10{,}93 \cdot ИР_i$$

$$d_2 = -15{,}65 - 13{,}44 \cdot МВ_i - 5{,}39 \cdot КП_i + 10{,}55 \cdot ИР_i$$

На основе полученных дискриминантных функций строятся классифицирующие функции вида:

$$d_{ik} = b_{k0} + b_{k1}x_{i1} + \ldots + b_{kp}x_{ip}, \quad k = \overline{1,3}$$

Объект относится к классу, у которого значение d_{ik} оказывается наибольшим. Коэффициенты классифицирующих функций удобнее вычислять по скалярным выражениям

$$b_{ki} = (n-3)\sum_{j=1}^{3}(w^{-1})_{ij}\,\overline{x}_{jk}$$

где b_{ki} - коэффициент для переменной i в выражении, соответствующему классу k,

$(w^{-1})_{ij}$ - обратный элемент внутригрупповой матрицы сумм попарных произведений W.

Постоянный член находится по формуле:

$$b_{k0} = -0.5 \cdot \sum_{j=1}^{3} b_{kj}\,\overline{x}_{jk}, \quad k = \overline{1,3}$$

В нашем случае получены три классифицирующие функции, которые отвечают соответственно за низкое, среднее и высокое качество математической подготовки:

$$d_{i1} = -31{,}08 + 8{,}09 \cdot МВ_i + 18{,}63 \cdot КП_i + 17{,}64 \cdot ИР_i$$

$$d_{i2} = -65{,}36 + 10.85 \cdot МВ_i + 27{,}58 \cdot КП_i + 26{,}21 \cdot ИР_i$$

$$d_{i3} = -109{,}34 + 13{,}83 \cdot МВ_i + 36{,}11 \cdot КП_i + 33{,}49 \cdot ИР_i$$

Для проверки точности классификации классифицирующие функции применяются к тем объектам, по которым они были получены. По доле правильно классифицированных объектов оценивается точность процедуры классифика-

ции. Результаты такой классификации представляют в виде классификационной матрицы, в которой указывается % правильно предсказанных наблюдений.

Классификационная матрица

Известный уровень качества	Предсказанный уровень качества (число/ процент)						
	1		2		3		всего
1	163	100%	0	0%	0	0%	163
2	0	0%	119	100%	0	0%	119
3	0	0%	0	0%	42	100%	42

Все наблюдения предсказаны точно, поэтому дискриминацию, осуществленную можно считать достоверной.

Таблица классификации наблюдений содержит значения классифицирующих функций, апостериорную вероятность, расстояние Махаланобиса.

Расстояние Махаланобиса используется для классификации наблюдений по классам. Объект относят к той группе, для которой расстояние D^2 наименьшее.

$$D^2 = (n-g)\sum\sum(w_{ij}^{-1})(x_i - \bar{x}_{ik})(x_j - \bar{x}_{jk}),$$

где D^2 – расстояние от точки X (классифицируемый объект) до центроида класса *i*.

Если предположить, что каждый объект должен принадлежать к одной из групп, то можно определить апостериорную верятность принадлежности *X* к *i*-ому классу (сумма апостериорных вероятностей - 1) по бейесовской схеме вычисления вероятностей:

$$\Pr(G_i / X) = \frac{\Pr(X / G_i)}{\sum \Pr(X / G_i)}$$

где $\Pr(X / G_i)$ - априорная вероятность принадлежности *X* к *i*-ому классу (доля элементов класса, находящихся дальше от центроида, чем *X).*

№	уровень качества	расстояние Махаланобиса			апостериорные вероятности			значения классифицирующих функций		
		1	2	3	1	2	3	1	2	3
1	1	4,515	31,894	80,218	1,000	0,000	0,000	13,28	-0,72	-25,91
2	1	3,144	13,392	47,145	0,996	0,004	0,000	30,92	25,49	7,58
3	1	3,653	7,614	32,937	0,908	0,092	0,000	40	37,71	24,03
4	3	38,970	7,894	1,588	0,000	0,108	0,892	76,27	91,50	93,63
5	3	62,856	20,160	5,225	0,000	0,002	0,998	85,82	106,86	113,29

6	2	9,747	2,088	18,783	0,029	0,971	0,000	49,55	53,07	43,69
7	2	9,747	2,088	18,783	0,029	0,971	0,000	49,55	53,07	43,69
8	2	30,574	11,118	13,441	0,000	0,900	0,099	66,72	76,14	73,97
9	1	3,134	7,871	35,683	0,936	0,064	0,000	39,01	36,34	21,41
10	1	3,653	7,614	32,937	0,908	0,092	0,000	40	37,71	24,03
11	**3**	62,856	20,160	5,225	0,000	0,002	0,998	85,82	106,86	113,29
12	**2**	13,200	0,030	10,783	0,002	0,996	0,002	57,64	63,92	57,52
13	**2**	9,747	2,088	18,783	0,029	0,971	0,000	49,55	53,07	43,69
14	**2**	21,002	7,056	15,320	0,001	0,993	0,006	58,63	65,29	60,14

Задание для самостоятельной работы

Имея достоверную информацию об обанкротившихся и не обанкротившихся в течении 5 лет предприятиях попытайтесь спрогнозировать поведение «молодых» фирм с помощью дискриминантного анализа.

Список литературы

1. Годин А.М. Статистические средние и другие величины и их применение в различных отраслях деятельности : учеб. пособие : рек. УМО/ А. М. Годин, В. Н. Русин, В. П. Соколин ; под общ. ред. А. М. Година. - 3-е изд.. - М.: Дашков и К, 2009. - 252 с.

2. Горшков, М. К. Прикладная социология : методология и методы : учеб. пособие : рек. УМО/ М.К.Горшков, Ф.Э.Шереги. - М.: Альфа-М: ИНФРА-М, 2009. - 415 с.

3. Добреньков, В. И. Методы социологического исследования : учеб. : доп. Мин. обр. РФ/ В.И. Добреньков, А.И. Кравченко. - М.: ИНФРА-М, 2009. - 768 с.

4. Дубов А.М. Многомерные статистические методы: учеб.: рек. Мин. обр. РФ/ А.М. Дубов, В.С. Мхитарян, Л.И. Трошин. -М: Финансы и статистика, 2000, 2003. - 352 с.

5. Ермолаев О.Ю. Математическая статистика для психологов: учеб. пособие: рек. РИС / О.Ю. Ермолаев. -3-е изд., испр. – М.: Изд-во Моск. психол. – соц. ин-та : Флинта, 2004. - 336 с.

6. Кузык, Б. Н. Прогнозирование, стратегическое планирование и национальное программирование: учеб.: рек. УМО/ Б. Н. Кузык, В. И. Кушлин, Ю. В. Яковец. - 2-е изд., перераб. и доп.. - М.: Экономика, 2008. - 576 с.

7. Куликов, Е. И. Прикладной статистический анализ : учеб. пособие : рек. УМО/ Е.И. Куликов. - 2-е изд., перераб. и доп.. - М.: Горячая линия - Телеком, 2008. - 463 с.

8. Фаддеев, М. А. Элементарная обработка результатов эксперимента : учеб. пособие/ М. А. Фаддеев. - СПб.: Лань, 2008. - 118 с.

9. Чесноков, С. В. Детерминационный анализ социально-экономических данных: [моногр.]/ С.В. Чесноков. - 2-е изд., испр. и доп.. - М.: ЛИБРОКОМ, 2009. - 168 с.

Наталья Николаевна Двоерядкина,
доцент кафедры общей математики и информатики АмГУ, кандидат педагогических наук

Алена Николаевна Киселева,
ст. преподаватель кафедры общей математики и информатики АмГУ

Татьяна Александровна Юрьева,
ст. преподаватель кафедры общей математики и информатики АмГУ

Многомерные математические методы моделирования социальных процессов: учебно-методическое пособие

Изд-во АмГУ.